TempoWordNet : une ressource lexicale pour l'extraction d'information temporelle

Mohammed Hasanuzzaman

Normandie Université

THESE

Pour obtenir le diplôme de doctorat

Spécialité Informatique

Préparée au sein de l'Université de Caen Normandie

TempoWordNet: A Lexical Resource for Temporal Information Retrieval

Présentée et soutenue par
Mohammed HASANUZZAMAN

Thèse soutenue publiquement le 19/01/2016 devant le jury composé de		
M. Patrice BELLOT	Pr., Aix-Marseille Université, France	Rapporteur
M. Adam JATOWT	Associate Professor, Kyoto University, Japan	Rapporteur
Mme Brigitte GRAU	Pr., ENSIIE, Évry, France	Examinatrice, présidente du jury
M. Stéphane FERRARI	MC HDR, Université de Caen Normandie, France	Directeur de thèse

Thèse dirigée par Stéphane FERRARI, laboratoire GREYC

ED SIMEM

UNICAEN UNIVERSITÉ CAEN NORMANDIE

GREYC

Présentation en français

Section rédigée par les encadrants, S. Ferrari, Y. Mathet et G. Dias.

Introduction

La compréhension de la temporalité d'objets ou d'informations est une clé pour raisonner sur la manière dont le monde évolue. Par nature, le monde est en constant changement, le temps est donc une de ses caractéristiques les plus importantes. Les événements, les changements, les circonstances qui demeurent sur une certaine période sont tous liés par leur ancrage dans le temps. Le temps permet d'ordonner événements et états, d'indiquer leur durée, de préciser leur début et fin.

Dans les dernières années, on peut noter un intérêt croissant pour les applications de Traitement Automatique des Langues (TAL) et de Recherche d'Information (RI) qui peuvent analyser la masse de données numériques disponibles, avec une demande croissante pour une prise en considération de la dimension temporelle. Pour la recherche d'information, face à la quantité d'informations disponibles, proposer un accès aux documents ou aux textes via leur dimension temporelle est particulièrement pertinent. Des applications comme Google News Timeline ou Inxight's TimeWall l'ont déjà en partie intégré. En extraction d'information, sous-domaine de la RI, il est particulièrement crucial d'associer la temporalité aux événements, aux faits extraits des textes. En résumé automatique, l'ordonancement chronologique des informations est essentiel pour les relater de manière cohérente. Pour les systèmes de question-réponse, la temporalité est centrale : il s'agit à la fois de reconnaître l'information temporelle de la question, parfois explicite mais bien souvent implicite, et de fournir une réponse qui lui corresponde.

L'expression du temps, de la temporalité, se fait de différentes manières en langue naturelle. Les expressions temporelles font appel à nombreuses notions, comme la position dans le temps, la durée, la fréquence, les relations d'interdépendance, etc. Le standard TIMEX2 permet d'en représenter la plupart, avec suffisamment de soup-

lesse pour intégrer l'imprécision de certaines ("longtemps", "plus tard"...). Ce travail porte plus particulièrement sur les notions de passé, présent et futur, associées à de nombreuses expressions temporelles.

L'étude menée porte plus spécifiquement sur la construction d'une ressource lexicale dans le but d'identifier les notions de passé, présent et futur, qu'elles soient exprimées explicitement dans des expresions temporelles ou qu'elles soient véhiculées différemment dans les textes. Afin d'en évaluer la pertinence et la qualité, cette ressource est exploitée dans les tâches de classification temporelle de phrases et de reconnaissance de l'information temporelle dans les requêtes (appelée par la suite reconnaissance de la *visée temporelle de requêtes*).

La thèse est structurée en six chapitres.

1. Une introduction en présente le cadre général, les motivations et les objectifs principaux.

2. Un état de l'art présente les notions théoriques et les méthodes pratiques dans le domaine de la RI, et plus particulièrement les travaux récents en recherche d'informations temporelles.

3. Le chapitre 3 se consacre à la présentation de *TempoWordNet*, un enrichissement de la ressource lexicale *WordNet* avec des informations temporelles associées à chacun de ses *synsets* (ensembles de synonymes), en s'inspirant des principes de SentiWordNet, mais dans un tout autre domaine. Après une présentation de travaux connexes, différentes méthodes sont étudiées pour construire cette ressource, exploitant un processus de propagation de l'information temporelle à travers les synsets ainsi que des méthodes de classification automatique. Un intérêt particulier est porté au problème de l'évaluation de la qualité de la ressource obtenue par le biais de ces méthodes. Une évaluation intrinsèque est menée par annotation manuelle et une évaluation extrinsèque par le biais de la classification temporelle de phrases.

4. Le chapitre 4 présente une application majeure de TempoWordNet : son usage pour la classification de la visée temporelle de requêtes. L'approche suivie s'appuie sur une technique d'apprentissage automatique exploitant les méthodes ensemblistes (*ensemble learning*), l'optimisation multi-objectifs, ici appliquée à la F-mesure harmonique. L'expérience, menée avec 28 classifieurs, permet d'améliorer l'état de l'art de manière conséquente.

5. Les premières versions de TempoWordNet ayant déjà été diffusées et utilisées par la communauté, le chapitre 5 présente une nouvelle méthode de construction de la ressource, dans le but d'en améliorer encore la qualité. Une approche itérative est mise en place, se fondant sur l'expansion des définitions des synsets d'une nouvelle version de TempoWordNet par les informations obtenues dans la précédente. Pour l'évaluation extrinsèque de cette nouvelle méthode, la classification temporelle de phrases est toujours utilisée, ainsi qu'une nouvelle application de classification temporelle de tweets.

6. La conclusion rappelle les principales contributions et avancées de ce travail avant de proposer quelques perspectives de recherche pour de futurs développements.

TempoWordNet

Le temps joue un rôle important pour toute information et peut être très utile dans les tâches de Traitement Automatique des Langues (TAL) et de Recherche d'Information (RI) telles que le tri chronologique d'événements, l'exploration de documents, la compréhension de requêtes ayant une visée temporelle ou encore la clasification. Avec un contenu numérique en croissance permanente provenant de différentes sources, présenter des informations pertinentes ancrées temporellement pour une meilleure satisfaction de l'utilisateur devient un objectif primordial des systèmes d'information actuels.

L'information temporelle est présente dans tout texte en langue naturelle, soit explicitement, c'est-à-dire sous la forme d'expressions temporelles, soit implicitement

sous forme de métadonnées ou par connotation dans certaines unités lexicales. La reconnaître et l'exploiter dans un but de recherche ou de présentation d'information sont des aspects importants qui peuvent améliorer significativement les fonctionnalités des outils de recherche. Dans ce travail, nous proposons de ce fait une ontologie temporelle, TempoWordNet, où une temporalité intrinsèque est associée au sens de chaque mot, répartie sur les quatre facettes suivantes : *atemporal* pour sans temporalité, *past* pour passé, *present* pour présent et *future* pour futur. Nous espérons ainsi pouvoir fournir une meilleure compréhension de la temporalité en langue, dont pourraient bénéficier tant les approches à grains fins (TAL) que celles à gros grains (RI).

Cette ressource est conçue comme une extension de WordNet. Le concept de temps est déjà présent dans la ressource initiale, mais ne concerne que quelques entrées dont le sens premier est li à la temporalité. Notre objectif est ici d'associer systématiquement une information temporelle à chaque entrée, pour rendre compte par exemple du rapport au futur d'un verbe tel que *to predict*, hyponyme de *to think*, sans lien avec le concept de temps dans WordNet. La méthode de construction de TempoWordNet s'appuie sur une analyse quantitative des définitions (ou *gloses*) associées aux synsets (ensembles de synonymes dans la ressource initiale WordNet), et sur l'utilisation de la représentation vectorielle résultante d'une entrée pour une classification semi-supervisée des synsets. Elle se fonde sur la sélection préalable d'un ensemble de synsets appelés germes, représentatifs d'une des catégories temporelles *past*, *present* ou *future*. L'idée sous-jacente est que les synsets temporels, en particulier les germes, devraient contenir une temporalité similaire dans leur définition. Le processus de classification est alors itéré sur des expansions répétées de la liste de synsets germes initiale. Nous étudions et expérimentons différents procédés d'expansion de cette liste : i) expansion lexico-sémantique ii) expansion probabiliste iii) expansion hybride.

Pour valider la ressource obtenue, nous menons différents types d'évaluations. Une évaluation intrinsèque est réalisée via la mesure d'accord inter-annotateurs sur

un extrait, montrant clairement une meilleure qualité de la ressource obtenue par expansion hybride. Nous proposons aussi une évaluation extrinsèque de TempoWordNet dans une tâche de *classification temporelle de phrases*. L'objectif principal de cette expérience est de vérifier notre hypothèse selon laquelle une ressource lexicale enrichie d'informations temporelles peut effectivement aider à classer des phrases selon les trois catégories temporelles passé, présent et futur.

Classification de la visées temporelle de requêtes

La recherche de documents sur le Web se fait par l'intermédiaire de requêtes que les utilisateurs soumettent à des moteurs de recherche. Constituées d'une phrase ou de quelques mots clés, ces requêtes comportent le plus souvent une facette temporelle, implicite ou explicite, qui vise à restreindre la recherche à une période temporelle spécifique. Bien comprendre ce ciblage temporel est alors essentiel pour ne pas noyer les résultats pertinents dans une foule de documents qui ne concernent pas la zone temporelle spécifiée. Cette spécification temporelle peut être de deux types : explicite, lorsqu'il est fait mention textuellement de la temporalité, comme dans "Jeux Olympiques 2012" (grâce à la mention "2012" spécifiant une année particulière), ou implicite, lorsqu'elle n'est pas marquée textuellement et doit alors être tirée du contexte, comme dans "prévisions météorologiques pour Paris" (où le contexte habituel d'une telle requête suggère les tous prochains jours). Ces deux catégories représenteraient respectivement 7% et 1% des requêtes observées, ce qui montre l'importance de leur bonne prise en compte. La tâche est difficile pour plusieurs raisons. La principale est le peu d'informations dont on dispose en entrée : la date de saisie de la requête, et les quelques mots soumis par l'utilisateur (le plus souvent 3 à 5 mots). Une seconde raison est la nature souvent ambiguë des mots, renforcée par le fait que les méthodes classiques de désambiguïsation sont souvent prises en défaut avec des phrases courtes.

Nous nous sommes intéressé au défi NTCIR-TQIC (Temporal Query Intent Classification, ou classification de la visée temporelle d'une requête) qui vise à classer

une requêtes parmi 4 catégories : Passé ("Cause de la mort de Mitterrand"), Présent ("Est-ce que Barcelone a gagné aujourd'hui ?"), Futur ("FIFA 2018") et Atemporel ("perdre du poids rapidement"). Il est en effet apparu que cette tâche était l'une des nombreuses applications envisageables de TempoWordNet. Nous avons utilisé la ressource d'apprentissage et de test fournie par NTCIR, qui consiste en 100 requêtes assorties de leur date d'émission et leur catégorie temporelle (soit 25 items par classe), et 300 requêtes pour les tests (soit 75 par classe) Il en découle une double difficulté : 1) corpus d'apprentissage extrêmement limité ; 2) les requêtes elles-mêmes étant très courtes, elle présentes intrinsèquement peu de caractéristiques permettant de les classer temporellement. La première difficulté est un problème classique d'apprentissage, tandis que nous tentons de pallier la seconde par une méthode d'expansion de la quantité d'information.

Notre méthode est la suivante : retenir le top K des snippets retournées par l'API Bing en réponse à la requête. L'idée maîtresse est que ces snippets amènent probablement des informations temporelles supplémentaires. Par exemple, "Cancer Michael Douglas" renvoie un snippet contenant explicitement "in 2010" et "in August 2010". Nous retenons alors chaque date, assortie d'un indice de confiance calculé par le web service GTE. Nous avons par ailleurs défini un ensemble d'attributs temporels indépendants, au nombre de 11, parmi lesquels l'indice de confiance des dates, la différence temporelle entre dates, le nombre de mots relatifs passé dans la requête, ou encore le nombre de snippets classés "Futur", etc. Tous ces attributs permettent d'atténuer le problème 2. Nous avons enfin utilisé une technique de combinaison de classifieurs, afin de pallier le problème 1 : plusieurs classifieurs restés imprécis en raison du manque de données peuvent se combiner en un agglomérat plus précis. Il s'agit de Multi Objective Optimization, ou Optimisation d'objectifs multiples. En particulier, nous avons utilisé le Non-dominated Sorting Genetic Algorithm (NSGA-II), en utilisant la solution qui maximise la F-Mesure, et en déployant le tout sur la plate-forme Weka.

Nos résultats dépassent largement les techniques de l'état de l'art, avec en particulier des gains de 10% pour le passé, 19% pour les événements récents, 9% pour le futur et 10% pour l'atemporal, soit un gain global de 16%.

Amélioration de la ressource

Les résultats obtenus ainsi que les retours des utilisateurs de TempoWordNet nous ont encouragé à développer une version plus fiable de la ressource. Pour terminer notre étude, nous proposons en conséquence une nouvelle stratégie de construction qui montre des améliorations notables par rapport aux versions précédentes de TempoWordNet. Elle s'appuie sur la construction de versions successives de TempoWordNet : pour chaque nouvelle version, les gloses sont d'abord enrichies par les synonymes des entrées temporelles qu'elles contiennent, selon la version précédente, puis la ressource est construite avec la stratégie de propagation hybride précédemment proposée.

Pour l'évaluation de cette nouvelle version, en plus des techniques précédentes (évaluation intrinsèque, extrinsèque via classification de phrases et de visée temporelle de requêtes), nous proposons une tâche de classification temporelle de tweets pour laquelle nous avons créé un corpus de référence via le crowdsourcing. L'ensemble des résultats montrent une qualité meilleure que les 3 versions précédentes : un Kappa amélioré de 0.2 et des gains de 5 à 6% en F-mesure sur les tâches externes.

Conclusion

Pour terminer, nous rappelons dans le dernier chapitre les différents apports de ce travail, puis nous proposons quelques perspectives de poursuite. Nous envisageons l'usage d'algorithmes de classification de graphes semi-supervisés pour tenter d'améliorer encore la qualité de la ressource. La mise en place d'une tâche de crowsourcing nous semble être une solution intéressante pour proposer un standard de référence pour l'évaluation intrinsèque, qui prendrait en considération la possibilité d'orientations temporelles multiples pour un même synset. Une version multilingue

de TempoWordNet est une piste de recherche importante à envisager. Enfin, si l'on considère les applications possibles, TempoWordNet semble particulièrement adapté pour engager le tri temporel des résultats de recherche de documents, mais pourrait aussi bénéficier au classement temporel de compte-rendus médicaux, à l'identification de sujets populaires dans les tweets traitant du futur, en lien avec les attentes ou craintes exprimées, ou encore s'intégrer dans une phase d'analyse temporelle d'un système de question-réponse.

Abstract

Time plays an important role in any information space and can be very useful in Natural Language Processing and Information Retrieval tasks such as temporal ordering of events, document exploration, temporal query understanding, and clustering. With ever growing digital content from diverse information sources, presenting relevant information anchored in time for better user satisfaction becomes more and more important for today's information systems.

Temporal information is available in every natural language text either explicitly, e.g., in the form of temporal expressions, or implicitly in the form of metadata, or connotatively in the form of lexical units. Recognizing such information and exploiting it for information retrieval and presentation purposes are important features that can significantly improve the functionality of search applications.

In this research, we introduce a temporal ontology namely TempoWordNet where word senses are associated to their intrinsic temporal dimensions: *atemporal*, *past*, *present,* and *future*. As such, we expect to provide a better understanding of time in language, which may benefit both fine-grained (NLP) and coarse-grained (IR) temporal studies.

The approach to construct TempoWordNet relies on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification. It is based on a set of manually

selected seed synsets that are marker of *past*, *present,* and *future* temporal categories . The underlying idea is that temporal synsets should embody temporality in their definition in a similar way. The classification process is iterated based on the repetitive semantic expansion of the initial seeds lists. We study and experiment different expansion processes: i) lexico-semantic expansion ii) probabilistic expansion iii) hybrid expansion.

Then, we intrinsically examine the quality of the resource. We also propose to experiment the usefulness of TempoWordNet based on an external task: *sentence temporal classification*. The main purpose of this experiment is to test our assumption that temporal knowledge-base can help to classify sentences into three different temporal categories: *past, present* and *future*.

Recognizing the underlying temporal intent of a query is a crucial step towards improving the performance of search engines. We examine how the temporal knowledge embedded in TempoWordNet can be useful in temporal query intent classification scenarios. Understanding temporal query intent is one of the many applications that can be developed using the resource.

Results and feedback of TempoWordNet use advocate for more reliable resource. At the end, we propose a strategy that shows steady improvements over the previous versions of TempoWordNet.

Acknowledgements

It would have not been possible to finish my dissertation without the guidance of Dr. Gaël Dias and support from my elder brother and wife.

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Stéphane Ferrari. It has been a pleasure working with him. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive. I would also like to thank him for his notable contribution to my personal and professional time at Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (GREYC).

I would like to thank Dr. Gaël Dias, who let me experience the research of temporal information processing for his excellent guidance, caring, patience, and continuous help and support.

I would like to thank Dr. Yann Mathet for guiding my research for the past several years and helping me to develop my background in temporal information processing and patiently corrected my writing.

Special thanks goes to the jury and other faculty members of GREYC for their advise and suggestions. I would like to thank Waseem, who as a good friend, was always willing to help and give his best suggestions. Many thanks to José, Paul, and other co-workers in the laboratory for helping me annotating data.

I would like to thank GREYC - CNRS UMR 6072 Laboratory for financially supporting my research.

I would also like to thank my parents, my sisters, elder brothers, sister-in-law, and brother-in-law. They were always supporting me and encouraging me with their best wishes.

Finally, I would like to thank my wife, Zarin Zahed. She was always there cheering me up and stood by me through the good times and bad.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Understanding the temporal property of object or information is key towards reasoning about how world changes. The world is ever-changing in its nature and time is the most important characteristic that occurs in this world. Things that happen and concern change (events), or circumstance that stay the same for a certain period of time (states) are related by their temporal reference. The concept of time comes into play to put events or states in sequence one after another, to indicate the duration of an event or state, and to specify when an event occurred and finished. Time seems to be utilized as a universal reference system to anchor, sequence, measure and compare the intervals occupied by events and states.

Recent years evidence unprecedented interest in Natural Language Processing (NLP) and Information Retrieval (IR) applications that can handle the wealth of electronic data available, with the demand of temporally aware systems becoming increasingly popular. This need is substantiated by the fact that most of the information available electronically has temporal dimension, in a sense that something that was true at some point of time could be untrue at another due to the constant changing nature of the world. Despite its ubiquitousness, consensus on how time could be formalized has historically been a difficult task. Moreover, incorporating time into automatic systems that can access the temporal dimension and extract temporal meaning of a text is more challenging than formalization of time.

1.1 Importance of Time in NLP and IR

The development and evaluation of temporal processing systems is not only an important research topic, but also a very practical challenge. As the amount of generated information increases so rapidly in the digital world, the notion of using time as another factor becomes more relevant to many applications such as information retrieval, automatic summarization, and question-answering etc.[Mani et al., 2004].

Information Retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Formally, IR is concerned with "finding material (usually documents) of an unstructured nature (usually texts) that satisfies an information need from within large collections (usually stored on computers)"[Manning et al., 2008]. With the rapid growth in digital information, both online and offline, the notion of time as a dimension through which information can be managed and explored becomes extremely relevant for IR. Incorporating temporal information could benefit numerous IR tasks such as clustering of search results according to various time dimensions (e.g. Google News Timeline [1]), or time-based browsing and exploration of search results using timelines (e.g. Inxight's TimeWall [2]).

Information Extraction (IE) is a sub-field of IR and is "the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts"[Cowie and Wilks, 2000]. Applying information extraction on text, is linked to the problem of text simplification in order to create a structured view of the information present in free text. The overall goal being to extract attributes of entities (e.g. a person's professional position), or relation between entities (e.g. the $employee\_of$ relation). In many situations, the extracted attributes and relations are legitimate only within certain temporal durations, as entities and their properties change over time. Therefore, it is very crucial to capture these temporal constraints to improve the overall performance of IE systems.

[1] Available online at: `http://news.google.com/`

[2] For more information: `http://www.inxightfedsys.com/products/sdks/tw/default.asp`

Automatic Summarization also attracts great deal of attention on the processing of temporal information. Automatic Summarization systems "take one or several texts and extract the ***most important*** information [...] from them"[Orasan, 2006]. Multi-document summarization would gain by introducing the relative order of events of news articles which overlap in their description of events. This temporal information is essential for assembling a chronologically coherent narrative from the events mentioned in diverse information sources. Automatic Summarization has many practical applications that include generating biographies, assisting journalists in preparing background information on breaking news, condensing clinical records and deriving the typical evolution of a disease, and so on.

Question Answering (QA) systems process large text collections to find "a short phrase or sentence that precisely answers a user's question"[Prager et al., 2006]. Temporal processing is the central part of a QA system intended to answer questions that explicitly request temporal information as their answer or questions that focus on an intrinsic time dependency. Examples are given below:

1. *Is Steve Jobs currently CEO of Apple Inc.?* ***(Intrinsic time dependency)***
2. *When did the French Revolution begin?* ***(Explicit temporal request)***
3. *Who was the youngest captain to lift the Champions League trophy?* ***(Intrinsic time dependency)***

Question number 2 can easily be answered if the document contains an explicit mention of the date the French Revolution began. However, in order to answer question number 1 & 3 correctly some advanced temporal analyzing techniques need to be applied to understand the semantics of sentences and paragraph.

Many applications would benefit from obtaining a precise temporal representation of a text, and with all the digital data available it is impossible to add temporal mark-up by hand, therefore the need for reliable temporal processing systems that can automatically understand the intrinsic temporal orientation of natural language text.

1.2 Temporal Expressions

Temporal expressions are natural language phrases that refer directly to time, giving information about when something happened, how long something lasted, or how often something occurred or denote calendar dates, times of day, periods of time, durations or sets of recurring times. Most temporal expressions in English play the syntactic role of circumstance adverbials that express the semantic role of time. Temporal expressions convey different types of time-related information: position, duration, frequency and relationship etc.[Biber et al., 1999].

A deep understanding of all types of temporal expressions found in natural language text is needed to be able to develop a automatic system that can approximate what a human does towards the portrayal of expressions that speak of to time. Towards this goal, different sets of annotation guidelines were laid for the representations of temporal expressions in text.

The TIMEX2, one of the widely used temporal annotation guidelines as well as other standard annotation schemes, differentiate between expressions capturing when something happened (**position in time**), how long something lasted (**duration**), or how repeatedly something occurs (**frequency**). Another significant distinction is made between expressions which can be normalized depending only on themselves alone and **underspecified, context dependent**, or **relative** temporal expressions. This type of expressions are known as **fully specified, context independent** or **absolute time**. **underspecified, context dependent**, or **relative** temporal expressions serve as an anchor to determine which specific time a **fully specified, context independent** or **absolute time** is meant. For instance, in Example 1, **seven o'clock March 15, 2015** represents the **fully specified** temporal category which embodies all the necessary information for its normalization. On the contrary, **next day** in Example 2 is an **underspecified** temporal expression where another **fully specified** temporal expression is required to anchor them on a timeline.

Example 1. *Ram returned to work* ***seven o'clock March 15, 2015****.*

Example 2. *Shyam started work the **next day**.*

TIMEX2 annotation guidelines also mention another type of expressions that do not indicate a specific time. For example:

Example 3. *It's important to find out why your **period** has gone on for **so long**.*

The phenomena we study in this thesis is related to another important class of temporal expressions covered by TIMEX2 annotation guidelines that mainly denote **past, present,** and **future** situation in natural language text. Examples are given below:

Example 4. *The **previous** game between Real Madrid and Chelsea washed out due to bad weather. (**past**)*

Example 5. *Subscribe for the **latest** gaming news. (**present**)*

Example 6. *Barack Obama **predicts** he could win a third term. (**future**)*

1.3 Objectives and Challenges

In this thesis, we investigate how to build a temporal domain ontology or knowledge-base (though ontology and knowledge-base differ in many aspects, we use these terms interchangeably) to identify temporal expressions and non-markable but time-related expressions that are marker of **past, present,** and **future** in natural language text. In particular, we propose to enrich an existing ontology namely WordNet with temporal information. Moreover, we study how to use this resource for some external tasks such as **sentence temporal classification** and **temporal query intent classification**.

The underlying idea for such study consists of a number of temporal processing operations and adoption of concepts that can be thought of as building blocks for developing temporally aware information retrieval and natural language processing

applications. The task is particularly challenging even for humans if they intend to formalize then in a knowledge-base understood by computers, despite the fact that they manage temporal information very naturally and efficiently during their everyday life. There are several explanations for this difficulty.

One explanation for the difficulty of automatically identifying temporal information that denote **past, present,** and **future** in natural language text is the fact that connotation of **past, present,** and **future** can be conveyed via a wide range of different mechanisms including tense, aspect, and lexical semantic knowledge [Pustejovsky, 2005]. These properties need to be correctly identified, interpreted, and combined to derive the appropriate temporal information.

Another challenge arises from the fact that temporal information is not always expressed explicitly, rather implicitly and require interpretations or inferences derived from world knowledge. For example, the sentences in Example 1 & 2 have similar syntax, but the events they describe are not in the same temporal order.

Example 1. *John fell. Mary pushed him.*

Example 2. *John fell. Mary asked for help.*

The temporal information in these examples is implicit, as the events described are neither anchored to precise points in time, nor specifically ordered with respect to neighboring events. To derive the correct temporal interpretation for these examples, one must rely on semantic content, knowledge of causation and knowledge of language use. Despite their structure and syntax being so similar, in the first example the event of *falling* is temporally after the event of *pushing*, while in the second example the event of *falling* precedes the event *asking*.

Most of the existing Information systems find it extremely difficult to understand semantic information of the type required to differentiate between the two examples above, and to infer correct temporal orientation in both examples. As a consequence,

the research community has concentrated mainly on the several mechanisms used by language to convey temporal information explicitly or implicitly.

With the growing population of digital knowledge over the web, finding information with effectiveness, efficiency and accuracy and represent them at concept label into an ontology becomes an increasingly challenging task, especially since most knowledge is contained in large collections of unstructured textual documents. Ordinary approaches are to acquire knowledge from the documents manually and then structure the knowledge at the conceptual level. However, conventional knowledge acquisition approaches are usually driven by humans, which means that they are labor-intensive, time-consuming and troublesome.

The development of domain ontology is important in building a list of vocabulary whereas the process of sharing and reusing this knowledge management can be accomplished easily. One of the significant challenges to develop the domain ontology usually concerns the lack of well defined semantics. Semantic are necessary to enable the entities such as applications or human to understand and determine which lexical to be stored as ontology. With the issues related to semantic representation, the quality and accuracy of the ontology will be a real concern. Moreover, to automatically solve the problem of ambiguity between the association terms is crucial towards building a domain ontology. The problem arises when different people may have different associations with one particular term. More precisely, WordNet is too fine-grained in its sense definitions that it does not distinguish between homographs (words that have the same spelling and different meanings) and polysemes (words that have related meanings). For example the term 'present' has 18 (eighteen) concepts in WordNet 3.0 [3]. Among these 18 (eighteen), 2 (two) concepts along with its definition are presented below:

1. ***Present, nowadays (synset: present.n.01)****: The period of time that is happening now.*

[3] https://wordnet.princeton.edu/wordnet/download/

2. ***Present (synset: present.n.02)****: Something presented as a gift.*

It is easy for human to understand that concept 1 i.e the sense (synset: present.n.01) associated with term 'present' has clear temporal connotation while concept 2 i.e. (synset: present.n.02) has no temporal indication in terms of past, present, or future. However, it is hard to automatically associate the correct concept to a term, especially when not enough context is available, to avoid any ambiguity in representing them inside the ontology.

1.4 Contributions

This dissertation presents a systematic investigation of how temporal information can be identified in natural language text. In particular, the research in this thesis proposes a framework to automatically time-tag words/tokens as *past, present, future,* and *atemporal* based on its intrinsic temporal orientation. We also show how Sentence Temporal Classification (STC) and Temporal Query Intent Classification (TQIC) tasks can be benefited from this time-tagging.

The study contributes to advances of temporal information research in NLP and IR mainly in three areas:

1. ***lexical resources for temporal processing.***

2. ***comparative evaluation.***

3. ***methodology to improve search application.***

To achieve this, a detailed study of the existing research on temporality both in NLP and IR is carried out. This study focuses on both linguistic and computational aspects of the field. Subsequently, TempoWordNet, a lexical resource for temporal information retrieval is introduced and exhaustively evaluated both intrinsically (manually) and extrinsically (corpus-driven). The main contributions of this research are presented below:

The **first main contribution** of this work is a temporal knowledge-base obtained by automatically enriching the WordNet which may contribute to the success of time related applications in NLP and IR. As expressed in [Strötgen and Gertz, 2013], time taggers usually contain pattern files with words and phrases, which are typically used to express temporal expressions in a given language (e.g. names of months). In fact, most temporal NLP tasks rely on a time-sensitive vocabulary. On the contrary, T-IR systems usually do not use information about time in language although they could benefit from it when facing the recurrent problem of missing explicit timexes.

We found that WordNet [Miller, 1995] is a good place to start to find time-sensitive concepts. Indeed, one can list a set of 21 temporal synsets by iteratively following the hyponymy relation from the concept of time (synset # 00028270) represented by the following gloss: *the continuum of experience in which events pass from the future through the present to the past.* In [Fellbaum, 1998b], author demonstrated that words related to tennis are not necessarily linked to the concept of tennis. Similarly, most temporal words are not under the concept of time. For example, concepts such as "prediction", "remember", "ancient", "fresh" clearly have a time dimension although they are not listed under the time subtree of WordNet. Therefore, we propose to enrich all WordNet synsets with their temporal dimensions and developed different TempoWordNets: TempoWordNet Lexical (TWnL), TempoWordNet Probabilistic (TWnP), and TempoWordnet Hybrid (TWnH) . The main contribution consists in tackling temporal expressions according to their semantic classification. The exclusive classification of temporal expressions guiding this work is also unique in the specialized literature.

The **second main contribution** of this thesis is that it performs a comparative and qualitative evaluation of the methodologies as well as the resources (TWnL, TWnP, TWnH) developed in the processes. The main motivation towards this comparative evaluation is to uncover the influence of different TempoWordNets i.e. TWnL, TWnP, and TWnH. In particular, we examined its usefulness based on an external task: Sentence temporal classification. The underlying idea is that a tempo-

ral knowledge-base can help to classify sentences into three different categories: *past*, *present* and *future*. The purpose of this qualitative evaluation is to inspect the merits of TempoWordNets intrinsically. In order to evaluate the time-tagged WordNets (TempoWordNets), we performed multi-rater inter-annotation process over statistically significant sample of randomly selected time-tagged synsets from different versions of TempoWordNet.

The **third main contribution** of this study is the development of novel resources, including:

- A corpus representing different temporal classes: *past*, *present* and *future* at sentence level. To accomplish this goal, we automatically selected a set of *past*, *present* and *future* sentences from the well-known SemEval-2007 corpus developed for task 15 [Verhagen et al., 2007] that involves identifying event-time and event-event temporal relations. This corpus is a version of TimeBank containing approximatively 2500 sentences with TimeML annotations. So, all sentences exclusively containing *past* (resp. *present*) expressions were marked as *past* (resp. *present*). As for *future*, all sentences containing *future* expressions combined or not with *present* timexes were tagged as *future*.

- A corpus of tweets conveying different temporal orientations (*past*, *present* and *future*). The corpus contains thousand (1000) temporal tweets. To prepare a set of unlabeled instances for annotation, tweets are collected using the Twitter streaming API[4]. At the first step, unlabeled tweets are classified as *past, recency, future* and *atemporal* by the Sentence Temporal Classifier (STC) obtained at the time of building TempoWordNet. Afterwards, to validate the classification, annotators of the CrowdFlower[5] platform were asked to choose from 4 (four) possible categories such as *past, present, future,* and *none of these* based on the underlying temporal orientation of tweets. An additional choice

[4] `https://dev.twitter.com/streaming/overview`. Last accessed on 02-05-2015
[5] `http://www.crowdflower.com/`

None of these is included to allow annotators to indicate ambiguous (tweets with unclear or multiple temporal orientations) and/or atemporal tweets.

The **fourth main contribution** of this dissertation is a framework to understand the temporal orientation behind a user's query. The underlying idea to look at how a search application can benefit from TempoWordNet. In oder to access this, a multi-objective optimization based ensemble learning paradigm is used to solve the problem of Temporal Query Intent Classification (TQIC). The overall idea of this task is to predict the temporal class (*past*, *recency*, *future*, *atemporal*) of a web search query given its issuing date.

The **fifth main contribution** of this research is a methodology to obtain a potentially more reliable TempoWordNet. Based on our findings at the time of building TempoWordNets, we introduced an iterative strategy that made use of some version of TempoWordNet to develop more precise TempoWordNets. Feedbacks and opinions/observations received from the community after public release of TempoWordNets motivated us to do so.

1.5 Structure of the Thesis

This thesis is structured into 6 (six) chapters which advance systematically: from necessary theoretical background in Information Retrieval (IR) to practical methods and approaches (Chapter 2), continuing with the original contributions in the areas of Temporal-Information Retrieval (Chapter 3); an application of TempoWordNet (Chapter 4); method to develop more reliable TempoWordNets (Chapter 5), and finishing with conclusions and future works (Chapter 6). Brief summary of each chapter is presented below:

Chapter 2 [15] discusses the fundamental concepts of information retrieval and related topics. In particular, it focuses on the foundations of temporal information retrieval approach. In this direction, we have presented a detailed discussion about time, its categorization, and how temporal information can be found in documents.

We also outlined the main concepts of temporal information extraction and the TimeML standard. This background chapter serves as introductory context for the rest of this thesis.

Chapter 3 [45] presents different strategies such as lexico-semantic, probabilistic, and hybrid towards automatic construction of temporal ontology namely TempoWordNet (TWnL, TWnP, TWnH), where each WordNet synset is time-tagged with 4 (four) dimensions: *past*, *present*, *future*, and *atemporal*. This chapter also covers the evaluation process of TempoWordNets. In order to evaluate the usefulness of TempoWordNets, different strategies are adopted. A sample of randomly selected TempoWordNets entries are manually evaluated using an inter-annotator agreement process. For automatic evaluation, we propose to evidence its usefulness based on sentence temporal classification. The underlying idea is that a temporal knowledge-base can help to classify sentences into three different categories: *past*, *present* and *future*.

Chapter 4 [75] looks at how a search application can benefit from TempoWordNet. In particular, we tackle the problem of identifying temporal intent of queries from a machine learning point of view. This chapter focuses on an ensemble learning solution, whose underlying idea is to reduce bias by combining multiple classifiers instead of relying on a single one. For our purpose, we made use of a set of features which can easily be extracted from TempoWordNet and different freely available resources.

Chapter 5 [97] presents an additional strategy to build more reliable and precise TempoWordNets. In this chapter, we introduce an iterative strategy that temporally extends glosses based on some version of TempoWordNet to obtain a potentially more reliable TempoWordNet. On top of sentence temporal classification to evaluate the resources, we propose an additional evaluation method i.e. tweet temporal classification.

Chapter 6 [109], last chapter of the thesis presents conclusions and potential future directions of research.

Chapter 2

Background

The objective of this chapter is to provide an introductory overview of Information Retrieval (IR) by providing basic technology, concepts, and models for modern IR, followed by Temporal Information Retrieval (T-IR). Then, we look at few significant research contributions in T-IR.

We start with an overview of the main ideas behind modern IR systems in Section 2.1. We cover briefly the foundation of Temporal Information Retrieval (T-IR) in Section 2.2. In this section we present a general introduction of time, temporal expressions and their categorization followed by an overview of Temporal Information Extraction. We conclude this section with an introduction to standard markup language for temporal expressions in natural language. Finally, Section 2.3 presents influential research works in T-IR.

2.1 General Overview of Information Retrieval

Information retrieval (IR) is a sub-field of computer science focused primarily on providing users the information resources relevant to an information need from a collection of information resources. IR deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The represen-

tation and organization of the information items should be such as to provide the users with easy access to information of their interest.

In terms of scope, the area has advanced from its early goals of indexing text and searching for useful information in a collection. Nowadays, research in IR includes modeling, Web search, text classification, systems architecture, user interfaces, data visualization, filtering, languages.

From the research perspective, the area could be studied from two rather distinct and complementary points of view: a computer-centered one and a human-centered one. The former is mainly deals with the building of efficient indexes, processing user queries with high performance, and developing ranking algorithms to improve the results. The later is mainly consists of studying the behavior of the user, of understanding their main needs, and of determining how such understanding affects the organization and operation of the retrieval system [BAEZA and Ribeiro-Neto, 2011].

2.1.1 Beginning of Information Retrieval

The practice of storing written information has started around 3000 BC, when the Sumerians assigned special areas to store clay tablets with cuneiform inscriptions [Singhal, 2001]. They developed special classifications techniques to identify every tablet and its content.

The usefulness of storing and retrieving written information became increasingly important over time, especially with inventions like paper and the printing press. After the invention of computers, people felt that they could be used for storing and retrieving information. The idea of using computers to search for relevant pieces of information was popularized in the article "*As We May Think*" by Vannevar Bush in 1945 [Bush and Think, 1945]. The first automated information retrieval systems were introduced in the 1950s. Several research works in the 1950s elaborated upon the basic idea of searching text with a computer. One of the most notable method

was introduced by H.P. Luhn in 1957, in which he proposed words as indexing units for documents and measuring word overlap as criterion for retrieval [Luhn, 1957].

Several key developments in the field happened in the 1960s. Most notable were the development of the *SMART* system by Gerard Salton [Salton, 1971]. Another groundbreaking work was done by Cyril Cleverdon and his group at the College of Aeronautics in Cranfield [Cleverdon, 1967]. The Cranfield tests developed an evaluation methodology for retrieval systems that is still in use by IR systems today. The *SMART* system, on the other hand, allowed researchers to experiment with ideas to improve search quality. A system for experimentation coupled with good evaluation methodology allowed rapid progress in the field, and paved way for many critical developments.

The 1970s and 1980s saw many developments built on the advances of the 1960s. Various models for doing document retrieval were developed and advances were made along all dimensions of the retrieval process. These new models/techniques were experimentally proven to be effective on small text collections (several thousand articles) available to researchers at the time. Though, Large-scale retrieval systems, such as the *Lockheed Dialog system*, came into use early in the 1970s. However, due to lack of availability of large text collections, the question whether these models and techniques would scale to larger corpora remained unanswered. This changed in 1992 with the beginning of *Text Retrieval Conference*, or *TREC* [Harman, 1993]. The US Department of Defense along with the *National Institute of Standards and Technology (NIST)*, cosponsored the *(TREC)* as part of the *TIPSTER* text program. The main goal of this conference was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection.

With large text collections available under *TREC*, many old techniques were modified, and many new techniques were developed (and are still being developed) to do effective retrieval over large collections. *TREC* has also branched IR into related but

important fields like retrieval of spoken information, non-English language retrieval, information filtering, user interactions with a retrieval system, and so on.

Despite its maturity, until recently, IR was seen as a narrow area of interest restricted mainly to librarians and information experts. Such a tendentious vision prevailed for many years, despite the rapid dissemination, among users of modern personal computers, of IR tools for multimedia and hypertext applications. Beginning of the 1990s evidenced rapid change in perceptions with the advent of the *World Wide Web (WWW)*.

Nowadays, the Web has become a universal repository of human knowledge and culture. Its success is based on the conception of a standard user interface which is always the same, no matter the computational environment used to run the interface, and which allows any user to create their own documents. As a result, millions of users have created billions of documents that compose the largest human repository of knowledge in history. An immediate consequence is that finding useful information on the Web is not always a simple task and usually requires posing a query to a search engine, i.e., running a search. And search is all about IR and its technologies. Thus, almost overnight, IR has gained a place with other technologies at the center of the stage.

2.1.2 Approaches to Information Retrieval

Broadly, there are two major approaches of IR technology and research: *Statistical* and *Semantic*. In statistical approaches, documents are retrieved and ranked highly that match the query most closely in terms of some statistical measures. While the later depends on the syntactic and semantic analysis of the text; in other words, they try to reproduce to some (perhaps modest) degree the understanding of the natural language text that a human user would provide. These two approaches are not in opposite and can be used together for various retrieval models.

Classical Approaches

Classical approaches are mostly dominated by statical methods. Popular categories are: boolean, extended boolean, vector space, probabilistic, and Inference Network Model.

- **Boolean/Extended boolean**: Early IR systems were boolean systems which allowed users to specify their information need using a complex combination of boolean ANDs, ORs and NOTs. Boolean systems have several shortcomings, e.g., there is no inherent notion of document ranking, and it is very hard for a user to form a good search request. Even though boolean systems usually return matching documents in some order, e.g., ordered by date, or some other document feature, relevance ranking is often not critical in a boolean system. Even though it has been shown by the research community that boolean systems are less effective than ranked retrieval systems, many power users still use boolean systems as they feel more in control of the retrieval process.

- **Vector Space Approach**: In the vector space approach text is represented by a vector of terms [Salton et al., 1975]. Words and phrases are typically considered as terms. If words are chosen as terms, then every word in the vocabulary becomes an independent dimension in a very high dimensional vector space. Any text can then be represented by a vector in this high dimensional space. If a term belongs to a text, it gets a non-zero value in the text-vector along the dimension corresponding to the term. Since any text contains a limited set of terms (the vocabulary can be millions of terms), most text vectors are very sparse. Most vector based systems operate in the positive quadrant of the vector space, i.e., no term is assigned a negative value.

  To assign a numeric score to a document for a query, the model measures the similarity between the query vector (since query is also just text and can be converted into a vector) and the document vector. If $\hat{D}$ is the document vector and $\hat{Q}$ is the query vector then the similarity of document D to query Q (or score of D for Q) can be represented as:

$$Sim\Big(\hat{D}, \hat{Q}\Big) = \sum_{t_i \in Q,D} \omega_{t_{iQ}} \cdot \omega_{t_{iD}} \tag{2.1}$$

where $\omega_{t_{iQ}}$ is the value of the ith component in the query vector $\hat{Q}$, and $\omega_{t_{iD}}$ is the ith component in the document vector $\hat{D}$.

- **Probabilistic Approach**: This IR approach is based on the general principle that documents in a collection should be ranked by decreasing probability of their relevance to a query. This is often called the *Probabilistic Ranking Principle (PRP)* [Robertson, 1977]. Since true probabilities are not available to an IR system, probabilistic IR models estimate the probability of relevance of documents for a query. This estimation is the key part of the model, and this is where most probabilistic models differ from one another. The initial idea of probabilistic retrieval was proposed by Maron and Kuhns in a paper published in 1960 [Maron and Kuhns, 1960]. Since then, many probabilistic models have been proposed, each based on a different probability estimation technique. Detailed mathematical background of all these models are out of scope of this thesis. However, the following description sums up the basis of these models.

  The probability of relevance R for document D is represented by $P(R \mid D)$. Considering the ranking condition is monotonic under log-odds transformation, document could be ranked by $\log \frac{P(R|D)}{P(\bar{R}|D)}$, where $P(\bar{R} \mid D)$ represents the non-relevant probability of document. It becomes $\log \frac{P(D|R)\cdot P(R)}{P(D|\bar{R})\cdot P(\bar{R})}$. Assuming that the prior probability of relevance, i.e. *P(R)*, is independent of the document under consideration and thus is constant accross documents, *P(R)* and $P(\bar{R})$ are just scaling factors for the final documents scores and can be removed from the above formulation. Above formulation could be further simplified to $\log \frac{P(D|R)}{P(D|\bar{R})}$.

Various probabilistic models came into the picture based on the assumption behind estimation of $P(D \mid R)$. One of the simplest form of this model, assuming that terms (words) are mutually independent, and $P(D \mid R)$ is transformed into the product of individual term probabilities, i.e., probability of presence/absence of a term in relevant/non-relevant documents:

$$P(D \mid R) = \prod_{t_i \in Q,D} P(t_i R) \cdot \prod_{t_j \in Q,\bar{D}} (1 - P(t_j \mid R)) \tag{2.2}$$

which uses probability of presence of a term t_i in relevant documents for all terms that are common to the query and the document, and the probability of absence of a term t_j from relevant documents for all terms that are present in the query and absent from the document. If p_i denotes $P(t_i \mid R)$, and q_i denotes $P(t_i \mid \bar{R})$, the ranking formula $\log\left(\frac{P(D|R)}{P(D|\bar{R})}\right)$ reduces to:

$$\log \frac{\prod_{t_i \in Q,D} p_i \cdot \prod_{t_j \in Q,\bar{D}} (1 - p_j)}{\prod_{t_i \in Q,D} q_i \cdot \prod_{t_j \in Q,\bar{D}} (1 - q_j)} \tag{2.3}$$

For a given query, we can add to this a constant $\log\left(\prod_{t_i \in Q} \frac{1-q_i}{1-p_i}\right)$ to transform the ranking formula to use only the terms present in a document:

$$\log \prod_{t_i \in Q,D} \frac{p_i \cdot (1 - q_i)}{q_i \cdot (1 - p_i)}$$

or

$$\sum_{t_i \in Q,D} \log \frac{p_i \cdot (1 - q_i)}{q_i \cdot (1 - p_i)}$$

Different assumptions for estimation of $p_i and q_i$ yield different document ranking functions. For example, [Croft and Harper, 1979] assumes that p_i is the same for all query terms and $\frac{p_i}{1-p_i}$ is a constant and can be ignored for ranking purposes. Okapi BM25 is another ranking function used by search engines to rank matching documents according to their relevance to a given search query.

It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by [Robertson et al., 1995].

- **Inference Network Approach**:

  In this technique, document retrieval is modeled as an inference process in an inference network [Turtle and Croft, 1989]. Most techniques used by IR systems can be implemented under this model. In the simplest implementation of this model, a document instantiates a term with a certain strength, and the credit from multiple terms is accumulated given a query to compute the equivalent of a numeric score for the document. From an operational perspective, the strength of instantiation of a term for a document can be considered as the weight of the term in the document, and document ranking in the simplest form of this model becomes similar to ranking in the vector space approaches and the probabilistic approaches described above. The strength of instantiation of a term for a document is not defined by the approach, and any formulation can be used.

Semantic Approaches

Semantic or *Natural Language Processing* (NLP) approaches to IR refer here to all methods based on the knowledge of *syntax* and/or *semantics* of the natural language in which document is written, or knowledge of the document genre, e.g., the application domains, to which the document refers. Such approaches refer as *semantic* approaches, in the sense that they attempt to understand the structure and meaning of textual documents.

These are broadly classifies into *phonological, morphological, lexical, syntactic, semantic, discourse, and pragmatic* according to the level of linguistic unit processed, and (correspondingly) the level and complexity of the processing required [Liddy, 1998]. The phonological level is the level of interpreting speech sounds, e.g., phonemes. It is mainly of interest in speech to text processing, rather than textual IR.

According to [Greengrass, 2000], semantics applies to (at least) five different levels of NLP. These levels are: the *morphological* level is concerned with analysis of the variant forms of a given word in terms of its components, e.g. prefixes, roots, and suffixes; the *lexical* level is concerned with analysis of structure and meaning at the purely word level; the *syntactic* level is the level at which the syntactic structure of sentences is determined, in terms of the parts of speech of the individual words. In practice, a single sentence can have many possible structures. Determining the correct structure from these alternatives requires knowledge at the higher levels (or statistics based on a training set); the *semantic* level is the level at which one tries to interpret meaning at the level of clauses, sentences, rather than just individual words; the *discourse* level is the level at which one tries to interpret the structure and meaning of larger units, e.g., paragraphs, whole documents, etc., in terms of words, phrases, clauses, and sentences; the *pragmatic* level is the level at which one applies external knowledge (that is, external to the document and original query). The knowledge employed at this level may include general knowledge of the world, knowledge specific to a given application domain, and knowledge about the user‘s needs, preferences, and goals in submitting a given query.

2.1.3 Evaluation in IR

Measuring the quality of a IR system can be investigated at different levels: (i) processing-time and space efficiency (ii) search-effectiveness of results (iii) system-satisfaction of the user. However, here we focus on evaluating retrieval effectiveness.

There are two widely used metrics for evaluating the outcome of a query: *precision* and *recall*. Many more measures for evaluating the performance of information retrieval systems have also been proposed. In general, measurement considers a collection of documents to be searched and a search query.

Metrics discussed here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query.

Assumption 1. There is a set of documents in the collection which is relevant to the search query as shown in Figure 2.1

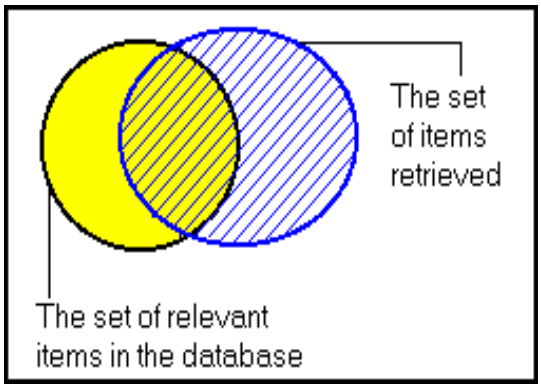

Figure 2.1: Set of relevant documents

Assumption 2. The actual retrieval set may not perfectly match the set of relevant documents as shown in Figure 2.2

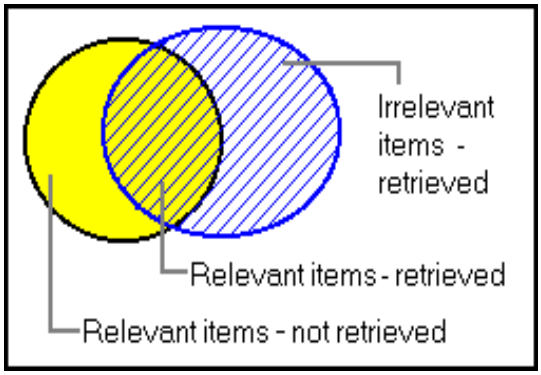

Figure 2.2: Relevant set of documents is not fully retrieved

Definition 2.1.1. RECALL is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection 2.3. It is usually expressed as a percentage.

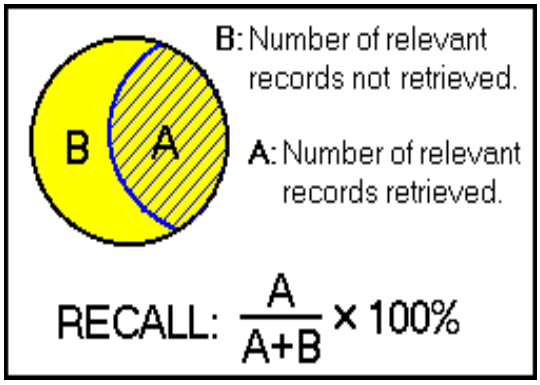

Figure 2.3: Recall

Definition 2.1.2. PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved 2.4. It is usually expressed as a percentage.

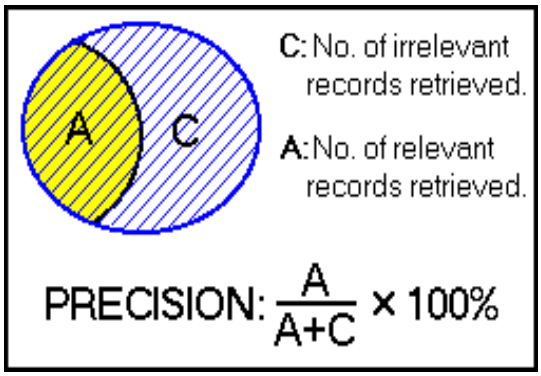

Figure 2.4: Precision

Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. In the Figure 2.5, the two lines may represent the performance of different search systems. While the exact slope of the curve may vary between systems, the general inverse relationship between recall and precision remains.

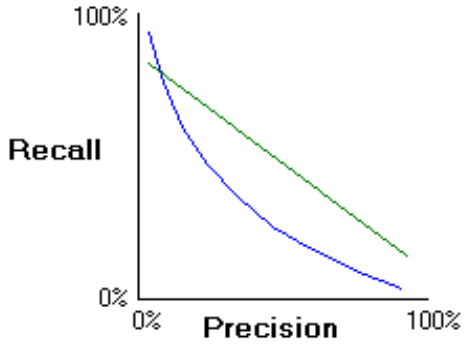

Figure 2.5: Performance of different information systems

There are various standard test collection and evaluation series exists that focus particularly on the evaluation for ad hoc information retrieval system.

The Cranfield collection[1], was the pioneering test collection in allowing precise quantitative measures of information retrieval effectiveness, but is nowadays too small for anything but the most elementary pilot experiments.

The Text Retrieval Conference (TREC)[2], co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, is a series of workshops that run a large IR test bed evaluation series since 1992. The main goal of TREC is to provide the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Within this framework, there have been many tracks over a range of different test collections, but the best known test collections are the ones used for the TREC Ad Hoc track during the first 8 TREC evaluations between 1992 and 1999. One can think of this as data sets that are used as benchmarks for evaluating different aspects of IR systems.

NII Test Collections for IR Systems (NTCIR)[3] project has built various test collections of similar sizes to the TREC collections, focusing on East Asian language and cross-language information retrieval, where queries are made in one language over a document collection containing documents in one or more other languages.

More recently, Cross Language Evaluation Forum (CLEF)[4] evaluation series has concentrated on European languages and cross-language information retrieval.

[1] http://ir.dcs.gla.ac.uk/resources/*test_collections*/cran/
[2] http://trec.nist.gov
[3] http://research.nii.ac.jp/ntcir/data/data-en.html
[4] http://www.clef-initiative.eu/web/clef-initiative/home

2.2 Foundation of Temporal Information Retrieval

With the rapid growth of digitised document resources, both on and off the web, and increased variety in types of document collections, future information systems will face growing difficulties in providing reliable, useful, and timely results. *Time* is a ubiquitous factor at many stages in the information-seeking process, with users having temporally relevant information needs, and collections having temporal properties at collection, document metadata, and document content levels.

Existing work in temporal information retrieval has aimed to account for time in document collections and relevance measurements. In contrast to classical information retrieval, temporal information retrieval aims to improve user experience by augmenting document relevance with temporal relevance.

Over the last few years, temporality has been gained an increasing importance within the field of information retrieval. Research on time and information retrieval covers a large number of topics [Alonso et al., 2011b, Campos et al., 2014a]. These include: the extraction of temporal expressions and events; temporal representation of documents, of collections and of queries; temporal retrieval models; temporal and event-based summarization; temporal text similarity; temporal query understanding; clustering of search results by time; temporality in ranking; visualization and design of temporal search interfaces; and so on.

This intense research is just in time to meet increasing demand for more intelligent processing of growing amounts of data. For example, social media data represents a sampling of all human discourse, and is temporally annotated with a document creation date. The historical (i.e., longitudinal) and emerging aspects of social media data are as yet relatively untapped [Derczynski et al., 2013].

The recognition of temporal information need and presentation of temporal information are very challenging problems. Expressions of time in documents are typically underspecified and vague ("I will see you later"-when? or, "As I was brushing my teeth." âĂŞ one needs to know how often teeth are brushed to guess when this was). Indeed, as humans experience time in the same way, temporality is not often expressed, instead remaining implicit. Further, understanding how to present data such that change, duration, order and other temporal aspects are clear is an area of research in IR. This difficulty of conveying temporality between system and user âĂŞ places a demand on builders of information systems to account for and model our understanding of time.

In the following sections, we present a general introduction to "*Time*". Afterwards, we study how time appears in text and their categorizations. An overview of Temporal Information Extraction and *TimeML* standard is presented in the subsequent sections.

2.2.1 What is Time?

Time has been a subject of interest of many researchers across various disciplines, particularly in physics, logic, philosophy, and art [van Benthem, 1991]. In physics, most important utilization of time is as a measure. For example, the time it takes an object to move from point A to point B. Time has also been at the center point of many advanced theorem in physics including Einstein's general relativity. Since ancient times, time has been studied by many philosophers. They have written about time and its impact on humankind in detail, most notably by Immanuel Kant and Martin Heidegger. Finally, in the arts time appears as a central object in paintings (Salvador), music (classic, pop), and literature (Jorge Luis Borges).

To set up the right context for our work, below contrastive definitions of time is studied. Formal definition of *"time"* and *"Temporal"* according to the Oxford Dictionaries are as follows:

Definition 2.2.1. TIME: The indefinite continued progress of existence and events in the past, present, and future regarded as a whole.

Definition 2.2.2. TIME: The continued progress of existence as affecting people and things.

Definition 2.2.3. TIME: Time or an amount of time as reckoned by a conventional standard.

Definition 2.2.4. TEMPORAL: The spatial and temporal dimensions of human interference in complex ecosystems.

2.2.2 Time in Text

Moving into the area of language, there are several ways of expressing time in a natural languages such as English, French, Hindi etc.. Without going into the theoretical or applied scientific study of *time* in language, we concentrate our study on the linguistic overview of *time*. A compressive study on language and time is excellently covered by the book written by Mani, Pustejovksy, and Gaizauskas [Mani et al., 2005].

The specific construction built into language for locating information in time is called *tense*. *Tense* is defined as *a set of forms taken by a verb to indicate the time (and sometimes also the continuance or completeness) of the action in relation to the time of the utterance* in the Oxford Dictionaries. The concept of tense in English is a method that we use to refer to time - past, present and future. There are three main tenses:

Definition 2.2.5. PRESENT TENSE: Things that are true when the words are spoken or written; or are generally true; or for some languages will be true in the future.

Definition 2.2.6. PAST TENSE: Things that were true before the words were spoken or written.

Definition 2.2.7. FUTURE TENSE: Things that will or might be true after the words are spoken or written.

Apart from tense, another linguistic mechanism called *grammatical aspect* indicates whether an *event* is considered as finished, completed, ongoing, upcoming or iterative. For example, *"Ram is going to the market"*. It is important to define the concept of *event*. An *event* could be defined as:

Definition 2.2.8. Event is something that takes place; an occurrence.

Definition 2.2.9. Event is a significant occurrence or happening.

Definition 2.2.10. Event is a social gathering or activity.

In corpora, an example of a temporal feature is temporal referring expressions that indicate times, durations or frequencies. Examples of such times are: "on Sunday", "March 15, 1983", "in the early 90sâĂİ, "two weeks", and "monthly".

Since language is ambiguous and evolves over time, there is more than just one way of representing the same time information. *"Christmas Day"* and *"December 25"* mean the same but are expressed differently: from a holiday perspective and from an entry in a calendar. The same time event can have different entries in a calendar depending on cultures. For example, *"Labor Day"* in USA is different from the rest of the world. Finally, the same point in time (same index in a calendar) can be expressed in different languages: *"Christmas"* in English and *"Noël"* in French.

2.2.3 Temporal Expressions

We now focus our study on the different types of temporal expression in text or associated with a document.

The first type of temporal information associated with a document is the *document metadata*. *Metadata* is the attribute of a document which express the date a document is created or last modified. Usually, such time related information of a

document is called as *document timestamp*. Document timestamps typically could be obtained at document collection crawling or indexing time and could be anchored in the *timeline*. A *timeline*, also known as a chronology, is a linear representation of events in the order in which they occurred.

The second type of temporal information is a little bit more involved as it relates to the linguistic analysis of the textual content of documents. A suitable approach to identify time in text data is information or named-entity extraction, with temporal entities being time-related concepts. Such concepts are represented in the document text as (not necessarily contiguous) sequences of tokens or words. In particular, temporal entities can be made explicit in the form of a temporal expression that correspond to a chronon in some timeline. A temporal expression is a portion of textual content that expresses some direct or inferred temporal information. These expressions include mainly dates ("2/08/2015") or prepositional phrases containing time expressions ("on Friday").

Temporal expressions are recognized by an entity extraction approach using a time-based linguistic analysis. Expressions can be mapped to temporal entities and terms defined in some temporal ontology.

According to the research carried out in NLP community, two types of temporal expressions are important to mention. In [Schilder, 2004], temporal expressions are classified as *time-denotating* and *event-denotating*. In this study, we are interested in *time-denotating* expressions, which can be explained as follows:

- **Explicit reference**: Expressions like "02/08/2015 that are entries in a calendar system. It also covers time expressions like "9 AM" or "Midnight" that denote a precise point in time.
- **Indexical reference**: Expressions that can be evaluated via an reference time. For example, "today" or "next Sunday" must be evaluated with respect to the document timestamp.

- **Vague reference**: Expressions that indicate only vague temporal information and it is difficult to place them on a timeline. For example, “in few weeksâĂİ, or “by Sunday the latest”.

The second classification proposed in [Pustejovsky, 2005] aims to cover different ways in which time is expressed in English Language. However, this classification is based on the *question-answering* in mind, where the system is able to retrieve answers to questions posed in natural language. For example, to answer a question about a person’s current affiliation, the system has to know all past affiliations of that person and select the newest one. A brief description of the different time expressions which fall under this classification are as follows:

- **Adverbial or prepositional phrases**: Expressions such as “in the 90s”, “Monday evening”, and “02/08/2015”.
- **Indexical**: Expression that by themselves do not specify a specific time and need an indexical anchor like “next Sunday”.
- **Indeterminate**: Expressions that cannot be interpreted as part of a timeline, because their begin and end points are not clear. For example:“in the summer”.
- **Durations**: Expressions that refer to quantities of time such as “for one week” and “a 5 hour journey”.
- **Anchored durations**: Expressions that specify the time of an event by making explicit the duration between the event and a time. For example: “one month from today”.
- **Set of times**: Expressions that groups distinct parts of the timeline. For example: “every Sunday” and “2 days per week”.

To conclude, depending on the application domain it may be necessary to provide a specific classification of how time information appears in text. Now, we take a closer look into [Schilder, 2004], and distinguish between explicit, implicit, and relative temporal expressions.

Explicit Temporal Expressions

Explicit temporal expressions describe chronons in some timeline, such as an exact date or year. It was first conceptualized during the Fifth Message Understanding Conference [Office, 1993]. For example, *"2015", "August 2015", and"15.08.2015"* for the *year, month, and day* granularity level respectively.

Implicit Temporal Expressions

Depending on the capabilities of the entity extraction approach and in particular its underlying time ontology, implicit temporal expression, such as names of holidays or events can be anchored in a timeline as well. For example, the temporal expression *"Christmas Day 2015"* found on a piece of text in a document can be mapped to the single expression *"December 25, 2015"*. However, it is also possible in some cases to map single temporal entity into more than one temporal expressions. For example, the token sequence *"Summer 2015"* can be mapped to a combination of month, week, and day chronons in three different timelines. In general, implicit temporal expressions require that at least a year chronon appears in the context of a named event.

Relative Temporal Expressions

Relative temporal expressions represent temporal entities that can only be anchored in a timeline in reference to another explicit or implicit, already anchored temporal expression. That is, their anchoring depends on a chosen point of time reference or narration. For example, the expression *"today"* alone cannot be anchored in any timeline. However, it can be anchored if the document is known to have a creation date as a reference. Then it is likely that the expression can be mapped to that date. There are many instances of implicit temporal expressions, such as the names of weekdays (e.g., "on Thursday") or months (e.g., "in July") or references to such points in time like *"next week"* or *"last Friday"*.

Relative temporal expressions may even include more ambiguous temporal information. Instances of these include phrases such as *"in a few weeks"* or *"some years ago"*. In general, there is less confidence in determining relative temporal expressions than in explicit or implicit expressions.

Although it might seem almost infeasible to detect and in particular anchor implicit temporal expressions, there have recently been significant advances in detecting and mapping instances of various types of implicit temporal information.

2.2.4 Temporal Information Extraction

The success of many NLP applications such as *text summarization, question answering* mainly depends on the identification of temporal information. In this section, we provide a brief overview of the important concepts, techniques, and tools for temporal information extraction.

[Mani et al., 2004], covers the state-of-the-art system in this field. The concept of time identification is first introduced as part of Named Entity (NE) tagging subtask within Message Understanding Conference (MUC), in particular MUC-6 [5]. In this task, participants were to identify (using SGML tags) named entities such as person name, organization name, date, time, currencies, and percentage. The quality of every participants system were measured in terms of *precision and recall* according to given standards. In MUC-6 date and time (and day) expressions were labeled using a TIMEX tag. In this task only absolute time expressions were required. In the following MUC-7 relative time expressions ("last summer") were also part of the task. The next section provides more details on tagging temporal expressions.

The main shortfall of the time identification tasks in MUC-6 and MUC7 is that there were no method to evaluate those temporal expressions to obtain a precise time. According to the MUC-7 TIMEX tagging guidelines, an expression like "yesterday" in a document about "August 15, 1998" would be tagged as a TIMEX of type DATE.

[5] http://www.itl.nist.gov/iad/894.01/tests/ie-er/er 99/er 99.htm

An application needs to know that "yesterday" is "August 14, 1998âĂİ. The TIMEX2 tagging guidelines address this issue by adding a calender value for every expression as an attribute of the tagged element.

It has been discussed and proposed in [Mani et al., 2004] that it is important to evaluate a *relational or indexical* temporal expression and return a formal calender time value. There is utility in separating the evaluation process in mapping the time expression and its evaluation. For example, *last Sunday* is normalized into the expression *sunday(predecessor(ts(d)))*, where *ts((d))* is the *metadata* which represents the document creation time. In the second phase, a final time is computed based on the value of *ts(d)*. The benefit of this approach is that the semantic interpretation of time is separated from anchoring.

2.2.5 TimeML

Because of the core task behind temporal expression identification is processing of language, it is important to follow some specific annotation guidelines. Example of such specifications are *TIMEX (MUC), TIMEX2, TIMEX3, and TimeML* [Pustejovsky et al., 2003, Pustejovsky et al., 2005a]. Earlier annotation tasks were performed using SGML tags with no specific Document Type Definition (DTD).

Most widely used and accepted temporal expression annotation specification in natural language is TimeML [Pustejovsky et al., 2005b]. TimeML language, a specification language for events and temporal expressions, and the relations held between them. It was first introduced in 2002 in a workshop called *Time and Event Recognition for Question Answering Systems (TERQAS)*[6], which mainly addressed to the issue of answering temporally based questions regarding events and entities in news articles. TimeML was further matured in the context of *TimeML Annotation Graphical Organizer workshop (TANGO)*[7] in 2003. In addition, TimeML has been consolidated as an international cross-language ISO standard (ISO WD

[6]http://www.timeml.org/site/terqas/index.html
[7]http://www.timeml.org/site/tango/index.html

24617-1:2007), and has been approved as the annotation language for *TempEval*, one of the tasks in the *SemEval International Workshop on Semantic Evaluations* [Verhagen et al., 2007, Verhagen et al., 2009]. TimeML has the following interesting properties:

- *Time stamping of events.*
- *Ordering events with respect to one another.*
- *Interpretation of partially determined expressions. Like any markup language, it has a DTD and Extensible Markup Language (XML) schema.*

TimeML has a number of tags described as follows: *EVENT* annotates events in a text. Any event that can be temporally anchored or ordered is captured in the tag; *SIGNAL* is used to annotate temporal functions words like "after", "during", etc.; There are three subtypes of *LINK*-temporal links (TLINK), subordination relationships (SLINK), and aspectual link (ALINK); The *TIMEX3* tag is primarily used to mark up temporal expressions, such as times, dates, durations, etc. The standard has four types to define time, date, duration, and set respectively.

An expression that receives the *TIME* type is one that refers to a time of the day, even if in a very indefinite way. For example: "five minutes to twelve", "half past noon, "9 AM, March 15, 1983", and "nine in the morning". The *DATE* type can be thought of as any extension that refers to a calendar time. For examples: "Monday, March 12, 2007âĂİ, "December 2015", "yesterday", or "last month. To distinguish between *TIME* and *DATE*, it is important to look at the granularity of the expression. If the granularity of the expression is smaller than a day, then the expression is of type *TIME*. An expression is a *DURATION* if it explicitly describes some extent of time. For example:"12 hours", "three months", or "7 days in December". Finally, the *SET* type is used for expressions that describe a set of regularly reoccurring times, for example, "twice a week" or "every Sunday". Below are examples of *TIME, DATE, DURATION*, and *SET*.

<TIMEX3 tid=“t1” type=“TIME” value=“1998-02-13T14:35:00” temporalFunction=“false” functionInDocument=“CREATION_TIME”>02/13/1998 14:35:00
</TIMEX3>
$< TIMEX3 tid = “t2” type = “DATE” value = “1998 - 08” temporalFunction = “true” functionInDocument = “NONE” anchorTimeID = “t41” > August < /TIMEX3 >$
<TIMEX3 tid=“t3” type=“DURATION” value=“P1M” temporalFunction=“false” functionInDocument=“NONE”>a month< /TIMEX3>
<TIMEX3 tid=“t4” type=“DATE” value=“PRESENT_REF” temporalFunction=“true” functionInDocument=“NONE” anchorTimeID=“t41”>current
$< /TIMEX3 >$
<TIMEX3 tid=“t5” type=“SET” value=“XXXX-WXX-1TNI” temporalFunction=“true” functionInDocument=“NONE” anchorTimeID=“t189” quant=“”
freq=“7D” > Monday$< /TIMEX3 >$

2.3 T-IR Research Works

This section presents some important research works in T-IR and its related subareas. The necessity of integrating temporal information was identified immediately after the emergence of information retrieval system at scale [Belkin and Croft, 1992]. *Internet Archive* founded by Brewster Kahle in April 1996, is one of the first organization to build temporally-aware information system to record the entire Internet [Kahle, 1997]. They collected the public materials on the Internet to construct a digital library. The collection includes all publicly accessible World Wide Web pages, the Gopher hierarchy, the Netnews bulletin board system, and downloadable software. The successful longitudinal system inspired work on other ways to access informa-

tion which includes the temporal dimension, especially for exploration and search purposes.

Later, research that integrated temporality into retrieval rankings become mature [Jensen and Snodgrass, 1999]. Today, major search engines have experimented bringing control of temporal search to the everyday user, with basic temporal refinement in their web search engine enabling filtering of results according to the publication time of the document.

Subsequently, standardization of the temporal semantics within documents developed, and formal definitions for "temporal expression" and "event" were prototyped. Research on temporal annotations is relatively new, and it is well covered in [Mani et al., 2005]. Identification of time depends heavily on the language and the corpora, so traditional information retrieval systems tend to fall short in terms temporal extraction. Based on the latest advances, new research is emerging for automatic assignment of document event-time periods and automatic tagging of news messages using entity extraction described in [Schilder, 2004, Schilder and Habel, 2001]. An example of guidelines for annotating temporal information is [Mani et al., 2001]. More sophisticated annotation schemes are presented in [Muller and Tannier, 2004]. An example of tools for annotation is TARSQI [Verhagen et al., 2005]. Recently, temporal taggers in different languages are available [Strötgen et al., 2014].

As time passes, different sub-areas of T-IR has been identified and approached, such as temporal query understanding, time-aware retrieval/ranking, temporal clustering, temporal search engines, future information retrieval, temporal snippets, temporal image retrieval etc. These lay the foundations for powerful analysis applications, with both general advances applicable across many areas and also tools and knowledge specific to certain domains.

Rosie Jones and Fernando Diaz pioneered the research on temporal query understanding in [Jones and Diaz, 2007], where they argued that timeline for a set of

documents returned in response to a query gives an indication of how documents relevant to that query are distributed in time. They proposed that by examining the timeline of a query result set allows to characterize both how temporally dependent the topic is, as well as how relevant the results are likely to be. They also showed that properties of the query result set timeline can help to predict the mean average precision of a query. Their findings also showed that meta-features associated with a query can be combined with text retrieval techniques to improve our understanding and treatment of text search on documents with timestamps. This dimensions is further pursued in [Dakka et al., 2008]. In this work, they proposed a general framework for handling time-sensitive queries and automatically identify the important time intervals that are likely to be of interest for a given query. In addition to that scoring techniques are built that seamlessly integrate the temporal aspect into the overall ranking mechanism for the news article data sets as well as real web data. Aggregated search by introducing news article in web search results is covered in [Diaz, 2009]. They addressed the issue of whether to integrate news content which changes over time or not for a given query. Their system adapts to news intent in two ways. First, to track development of and interest in topics by inspecting the dynamics of the news collection and query volume. Second, to quickly recover from system errors click feedback can be used. For doing this, several click-based metrics were defined which allow a system to be monitored and tuned without annotator effort. The time of queries using temporal language model for the queries that comprise only keywords, and their relevant documents are associated to particular time periods not given by the queries is particularly tacked in [Kanhabua and Nørvåg, 2010]. Different approaches are exercised to automatically determine the temporal nature of queries especially the implicit one is performed in [Campos et al., 2011b]. They exploited web snippet and query log for this task. Some of the notable research works in this research area include [Shokouhi, 2011, Campos et al., 2012b, Campos et al., 2012c, Zhang et al., 2010, König et al., 2009].

There is a growing body of scientific works which exploit the idea of introducing time in retrieval model. [Li and Croft, 2003] is one of the initial efforts which

explored the relation between time and relevancy. First some queries are identified that favors very recent documents. Then a time-based language model approach is adopted to retrieve documents for these queries by incorporating time into both query-likelihood models and relevance models. These models were used for experiments comparing time-based language models to heuristic techniques for incorporating document recency in the ranking. Temporally adaptive content-based relevance ranking algorithm that explicitly takes into account the temporal behavior of the underlying statistical properties of the documents in the form of a statistical topic model is proposed in [Perkiö et al., 2005]. Method for ranking search results for year qualified temporal queries are presented in [Zhang et al., 2009]. The method adjusts the retrieval scores of a base ranking function according to time-stamps of web documents so that the freshest documents are ranked higher. Another interesting work which uses the revision history of a document (e.g., the edit history of a page in Wikipedia) to redefine term frequency - a key indicator of document topic/relevance for many retrieval models and text processing tasks is [Aji et al., 2010]. Few other relevant works in this area are [Dong et al., 2010b, Dong et al., 2010a, Costa et al., 2014, Kanhabua and Nørvåg, 2012, Chang et al., 2012].

Clustering of search results is an important feature in many of today's information retrieval applications. The notion of hit list clustering appears in Web search engines and enterprise search engines as a mechanism that allows users to further explore the coverage of a query. Temporal attributes of documents content such as a date and time token or as a temporal reference in a sentence are taken into account for constructing and presenting clusters [Alonso and Gertz, 2006a, Alonso et al., 2009b]. Other contributions in this area are [Mori et al., 2006, Campos et al., 2009, Campos et al., 2012c].

As search applications keep gathering new and diverse information sources, presenting relevant information anchored in time becomes particularly important for both expert users, e.g.,historians, librarians, and journalists, as well as a general user searching for information needs in old versions of web pages. [Alonso et al., 2007a]

presents an exploratory search interface that uses timelines to present and explore search results. The timeline construction is based on a clustering algorithm that uses temporal expressions extracted from documents and anchoring these expressions in a timeline [Alonso and Gertz, 2006b]. [Jin et al., 2008] proposed a temporal search engine supporting content time retrieval for Web pages. The main purpose of this work is to support the Web search on temporal information embedded in Web pages. It is based on a unified temporal ontology of Web pages. The time in this ontology is defined to denote the most appropriate time describing the content of a Web page. Another significant work in this area is an on-demand search engine namely ChronoSeeker [Kawai et al., 2010]. The main goal behind this search engine is to collect as many future/past events as possible relevant to user's query by obtaining various future scenarios considering both predictions and histories. Another application called Time Explorer is designed for analyzing how news changes over time [Matthews et al., 2010]. It has been built as part of the European project Living-Knowledge [8]. The goal is to provide tools that allow exploring knowledge from all points of view and crucially to see how knowledge evolves over time.

Humans have always wanted to know their future, resorting from religious texts and astrology, to fortune tellers. Although we cannot know the future, a lot can be inferred about it by gathering huge amount of future information from web. Ricardo BaezaÂŋYates introduced the idea of *future retrieval* in *ACM SIGIR 2005 Workshop on Mathematical/Formal Methods in Information Retrieval* [Baeza-Yates, 2005]. The main motivation is to use news information to obtain future possible events and then search events related to our current (or future) information needs. In this work, *future retrieval* system is composed of a simple probability model for future events, a model based on a set of time segments and a generic simple ranking extension to any IR model. This idea of *future retrieval* is further pursued in [Jatowt et al., 2009]. It proposed to automatically generate summaries of future events related to queries using data obtained from news archive collections or from the Web. Two methods, explicit and implicit future-related information detection

[8] http://livingknowledge-project.eu/

are mainly covered in this work. Retrieval of explicit future related information is based on analyzing the context of future temporal expressions in documents, while the latter relies on detecting periodical patterns in historical document collections. Future-related information which is grounded in time, that is, the information on forthcoming events whose expected occurrence dates are already known is examined in [Jatowt et al., 2010]. Future retrieval is examined from different perspective by the following articles: [Dias et al., 2011, Kanhabua et al., 2011a, Kanazawa et al., 2011, Jatowt and Au Yeung, 2011, Campos et al., 2011a, Radinsky and Horvitz, 2013a].

Snippets available in Web search engines present a couple of lines with highlighted keywords and some context. There are no complete sentences, only bits of text that one must mentally construct as a sentence. One can argue that a document snippet that leverages temporal information would be an interesting alternative for some document search and exploration tasks. Intuitively, it makes sense to include time in a snippet. The idea of temporal snippet is introduced in [Alonso et al., 2009a] and subsequently progressed by the works carried out in [Alonso et al., 2011a, Svore et al., 2012].

Temporal Web Image Retrieval can be defined as the process that retrieves sets of Web images with their temporal dimension from explicit or implicit temporal text queries [Dias et al., 2012]. Inspired by recently emerging interests on query dynamics in information retrieval research, time-sensitive image retrieval algorithm can infer users implicit search intent better and provide more engaging and diverse search results according to temporal trends of Web user photos [Kim and Xing, 2013]. It modeled observed image streams as instances of multivariate point processes represented by several different descriptors, and develop a regularized multi-task regression framework that automatically selects and learns stochastic parametric models to solve the relations between image occurrence probabilities and various temporal factors that influence them. Automatically estimating the age of historical color photographs is performed in [Palermo et al., 2012, Martin et al., 2014].

Most recently, focus has turned to our interactions with temporality, including our behaviour and how to present information that has temporal parts. Graphical representations of temporal information are hard to create, confused by imperfect metaphors and underspecification [Plaisant et al., 1998, Verhagen, 2007]. In terms of visual information access, *Google NGram Viewer* has been released as basic tool for mining the rise and fall words used in five million books over selected years. *MIT* has developed *SIMILE Timeline Visualisation*, a Web widget prototype for visualising temporal data. Organizing, searching over and mining past information in terms of events has proven a difficult and interesting challenge, and making headway is yielding interesting results [Strötgen and Gertz, 2012, Talukdar et al., 2012]. Commercial products have focused not only on historical search, but also search over future information, such as Recorded Future and Yahoo!âĂŹs Time Explorer application [Matthews et al., 2010]. This direction is fueled by current research, such as Radinsky and HorvitzâĂŹs system [Radinsky and Horvitz, 2013b]. Demanding as these challenges are, advances in being temporally aware while presenting, mining and analysing data have led to extremely powerful results.

2.4 Summary

We presented the fundamental concepts of information retrieval and related topics. In particular, we have discussed the foundations of temporal information retrieval approach. In this direction, we have portrayed a detailed discussion of time, its categorization, and how temporal information can be found in documents. We also outlined the main concepts of temporal information extraction and the TimeML standard. This background chapter serves as introductory context for this research.

Chapter 3

TempoWordNet

In this chapter, we discuss in detail the building of a temporal ontology, which may contribute to the success of time-related applications. Precisely, we introduce the *TempoWordNet*, a lexical knowledge base where each synset of WordNet is augmented with its intrinsic temporal value that serves as the central point for the rest of the thesis.

This chapter is organized as follows. The next section motivates our approach and sets the objectives. Section 4.2 discusses related work. Afterwards, Section 3.3 presents various ways to build the temporal ontology. Precisely, we discuss the lexico-semantic, probabilistic, and hybrid expansion strategies followed to build different versions of TempoWordNet. Finally, Section 5.3 explains the experiments carried out to access the usefulness and quality of the different versions of TempoWordNet.

3.1 Motivation and Objectives

Time plays a crucial role in any information search, and therefore has been exploited in several information retrieval tasks such as information extraction, topic detection and tracking, question answering, summarization, and clustering. Since its inception, all the temporal information embedded in documents have not been fully utilized by IR applications for better user satisfaction and additional search features.

In natural language texts, different types of temporal information may be associated. For example, most evident type of temporal information associated with a document is metadata information i.e. its creation time or the modification time. This type of information can easily be identified, accessed, and used for several tasks, for example time-aware search, temporal clustering, temporal ranking etc. It is important to mention that the meta information is only valuable in some specific contexts. For example, only *document creation time* could be useful for news domain. However, apart from the news domain or even news domain itself, to consider only the *document creation time* means to overlook lot of other temporal information associated with a document. It is ubiquitous because lot of latent temporal information is available inside document's text. Assume a news document reporting about some future event. Considering only the meta data information i.e. document creation time will lead to spurious outcome about the event time. But to make use of such latent temporal information, usually temporal taggers are applied to extract and normalize temporal expressions contained in documents.

Time can be expressed in countless manners in natural language texts. Therefore, it is infeasible to catch all the temporal expressions with State-of-the-art temporal taggers such as SUTime [Chang and Manning, 2012, Strötgen and Gertz, 2013]. Temporal taggers mainly rely on regular expressions expressed by surface tokens and their Part-of-Speech (POS) categories and hardly bridge the gap when no mention of temporal expressions are available. Temporal expressions can be grouped into two main categories [Alonso et al., 2007b]: Absolute temporal expressions divided into Explicit temporal expressions (e.g., March 2, 2015) or Implicit ones (e.g., May Day 2015), and Relative temporal expressions (e.g., last year). However, thousands of words exists such as *'past', 'present', 'current', 'future', 'forecast', 'upcoming' etc.* that clearly posses a time dimension. For example, sentence like *'Your iPhone 4 or 5 could feel like new again with upcoming iOS9'* is clearly carrying future connotation. The word "upcoming " has a clear future connotation though ignored by most of the existing temporal tagging methodologies.

Most of the temporal NLP tasks rely on a time-sensitive vocabulary. On the contrary, T-IR systems usually do not use information about time in language although they could benefit from it when facing the recurrent problem of missing explicit timexes. To overcome this problem we intend to build a *temporal ontology* which may benefit both fine-grained (NLP) and coarse-grained (IR) temporal studies.

3.2 Related Work

In this section we focus on some works that are relevant to this research. In particular, works related to the temporal phenomenon in NLP and IR domains. We will also concentrate on the automatic construction of ontologies.

A great deal of works have been proposed in temporal NLP. Most recent studies have been developed in the context of the TempEval evaluation contests which were initiated by [Verhagen et al., 2007]. TempEval was initially divided into three challenges: (task A) identifying temporal relations between events and time expressions, (task B) identifying temporal relations between events and the document creation time and (task C) identifying the temporal relations between contiguous pairs of matrix verbs. In TempEval-2 [Pustejovsky and Verhagen, 2009], the best performing systems were based on conditional random fields mixed with parsing methodologies [UzZaman and Allen, 2010]. More recently, in TempEval-3 [UzZaman et al., 2013], new systems have been performing at high level of performance for all three tasks such as the rule-based multilingual temporal tagger Heideltime [Strötgen and Gertz, 2010].

Various research studies have also been proposed in temporal IR. The work of [Baeza-Yates, 2005] defines the foundations of T-IR. Then different researches have been tackled in several parallel topics, such as user query understanding [Metzler et al., 2009], temporal web snippets generation [Alonso et al., 2007a], temporal ranking of documents [Kanhabua et al., 2011b], temporal clustering

[Alonso et al., 2009c], future retrieval [Radinsky and Horvitz, 2013b] or temporal image retrieval [Dias et al., 2012].

As expressed in [Strötgen and Gertz, 2013], time taggers usually contain pattern files with words and phrases, which are typically used to express temporal expressions in a given language (e.g. names of months). In fact, most temporal NLP tasks rely on a time-sensitive vocabulary. On the contrary, T-IR systems usually do not use information about time in language although they could benefit from it when facing the recurrent problem of missing explicit timexes.

An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest [Gruber, 1995], where formal implies that the ontology should be machine-readable and the domain can be any that is shared by a group or community. Much of current research into ontologies focuses on issues related to ontology construction and updating. In general, there are two main approaches to ontology building: (i) manual construction of an ontology from scratch, and (ii) semi-automatic construction using tools or software with human intervention. Manual ontology construction is costly, time-consuming, error-prone, and inflexible to change. It is expected that an automated ontology learning process will result in more effective and more efficient ontology construction and also be able to create ontologies that better match a specific domain [Maedche, 2002]. Whereas semi-automatic generation of ontologies will substantially decrease the amount of human effort required in the process [Hotho et al., 2005, Luong et al., 2009, Omelayenko, 2001].

Semi-automatic construction of ontologies has been the central point of attraction among ontology engineers and domain experts since many years. Ontology learning uses methods from a diverse spectrum of fields such as machine learning, knowledge acquisition, natural language processing, information retrieval, artificial intelligence, reasoning, and database management [Shamsfard and Abdollahzadeh Barforoush, 2003]. A thorough summary of several ontology learning projects that are concerned with knowledge acquisition from a

variety of sources such as text documents, dictionaries, knowledge bases, relational schemas, semi-structured data, etc. is covered in [Gómez-Pérez et al., 2003].

An important task of ontology creation is to enrich the vocabulary for domain ontologies using different sources of information. Rather than working with domain-relevant documents from which vocabulary can be extracted, some ontology construction techniques exploit specific online vocabulary resources. WordNet, an online lexical database covering many domains, has been widely used as a source from which to mine new vocabulary for ontology enrichment. First notable work towards this direction is the OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet, that aim at achieving a formal specification of WordNet [Gangemi et al., 2003]. Within this project, they developed a hybrid bottom-up top-down methodology to automatically extract association relations from WordNet, and to interpret those associations in terms of a set of conceptual relations.

Another inspiring idea to use WordNet to enrich vocabulary for ontology domain is proposed by Luong et al., [Luong et al., 2009]. Their method relies upon the lexical expansion inside WordNet network to accurately extract new vocabulary for an ontology for any domain covered by WordNet. Few more research works that aim to extract words from WordNet's lexical database to enrich ontology vocabularies: Lexical Enrichment of a Human Anatomy Ontology using WordNet [Reiter and Buitelaar, 2008]; efficient text mining approaches combined with semantic information from WordNet [Speretta and Gauch, 2008]. In [Reiter and Buitelaar, 2008], Reiter et al. describe an approach that combines the Foundational Model of Anatomy with WordNet by using an algorithm for domain-specific word sense disambiguation. Speretta et al. present an approach that can enrich the vocabulary of each concept with words mined from a set of documents combined with semantic information from WordNet [Speretta and Gauch, 2008].

We also found that WordNet is a good place to start to find time-sensitive concepts. Indeed, one can list a set of 21 temporal synsets by iteratively following the

hyponymy relation from the concept of time (synset # 00028270) represented by the following gloss: *the continuum of experience in which events pass from the future through the present to the past.* However, likewise the tennis problem evidenced in [Fellbaum, 1998b], most temporal words are not under the concept of time. For example, concepts such as "prediction", "remember", "ancient", "fresh" clearly have a time dimension although they are not listed under the time subtree of WordNet. Therefore, based on the initial ideas of [Moens and Steedman, 1987] on temporal ontologies and inspired by SentiWordNet [Esuli and Sebastiani, 2006], we propose to enrich all WordNet synsets with their temporal dimensions. Moens and Steedman [Moens and Steedman, 1987], present an approach to build temporal ontology based on the notions like tense, aspect, and certain temporal adverbials, and a theory of their use in defining the temporal relations of events rather than on purely temporal primitives. It assumes that the temporal categories of tense, aspect, aspectual adverbials and of propositions themselves refer to a mental representation of events that is structured on other than purely temporal principles, and to which the notion of a nucleus or contingently related sequence of preparatory process, goal event and consequent state is central. This work also advocates that a principled end unified semantics of natural language categories like tense, aspect and aspectual/temporal adverbials requires an ontology based on contingency rather than temporality.

Kamps et al. [Kamps et al., 2004] use WordNet to construct a network by connecting pairs of synonymous words. The semantic orientation of a word is decided by its shortest paths to two seed words "goodâĂİ and "bad" representing positive and negative orientations. Motivated by the idea of [Kamps et al., 2004], Esuli and Sebastiani [Esuli and Sebastiani, 2006] used text classifier to classify orientations of words. Their method determines the orientation of words based on glosses in an online glossary or dictionary. The classifier is trained on glosses of selected seed words and is then applied to classify gloss of an unknown word to categorize the word as positive or negative. In the following sections, we propose a similar method, as well as new ones, to build temporal lexical resources.

3.3 Building TempoWordNet

Temporal aspects are fundamental to natural language interpretation. One could argue that there is no sentence or utterance the interpretation of which does not involve temporal aspects. Therefore, our approach to ontology-based temporal connotation of natural language has to account for these temporal aspects. For this purpose, we need an ontology or logical vocabulary that allows us to represent the temporal connotation of the meaning of a sentence, utterance or discourse. Instead of building a new ontology, we build a temporal ontology namely TempoWordNet based on the WordNet [Fellbaum, 1998a, Miller, 1995]. Before proceeding further, we will briefly discuss about the WordNet.

Wordnet is an electronic resource of synonyms, thesaurus, lexical database, taxonomy of concepts based on psycholinguistics studies. With its broad coverage and a design that is useful for a range of NLP and IR applications, this resource has found wide general acceptance. WordNet is perhaps the most important and widely used lexical resource for NLP applications up to now. The initiative of building WordNet in the mid-1980s was inspired by current theories of human semantic organization [Collins and Quillian, 1969]. People have knowledge of about tens of thousands of concepts, and the words expressing these concepts must be stored and retrieved in an efficient and economic way. A semantic network such as WordNet is an attempt to model one way in which concepts and words could be organized.

The basic unit of WordNet is a set of cognitively equivalent synonyms, or synset. Wordnet is mainly composed of nouns, verbs, adjectives, and adverbs synsets. Examples of a noun, verb, adjective, and adverb synset are {*vacation, holiday*}, {*vacation, holiday*}, and {*good*}, and {*well, good*} respectively. Each synset represents a concept, and each member of a synset encodes the same concept. In other words, synset members are interchangeable in many contexts without changing the truth value of the context. Each synset also includes a definition, or 'gloss' and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with

several distinct meanings are represented in as many distinct synsets. Thus, each form-meaning pair in WordNet is unique.

The main relation among words in WordNet is synonymy, as between the words *shut and close* or *car and automobile*. Synonyms relates words that denote the same concept and are interchangeable in many contexts and are grouped into unordered sets (synsets).The current version of WordNet [1] contains over 117,000 synsets that are organized into a large semantic network. The synsets are interlinked by means of different bidirectional conceptual-semantic and lexical relations. Following is the list of relations available in WordNet:

- **Synonymy**: Relation binding two equivalent or close concepts (*frail /fragile*). It is a symmetrical relation.
- **Antonymy**: Relation binding two opposite concepts (*small /large*). This relation is symmetrical.
- **Hyperonymy**: Relation binding a $concept_{-1}$ to a more general $concept_{-2}$ (tulip /flower).
- **Hyponymy**: Relation binding a $concept_{-1}$ to a more specific $concept_{-2}$. It is the reciprocal of hyperonymy.
- **Meronymy**: Relation binding a $concept_{-1}$ to a $concept_{-2}$ which is one of its parts (*flower/petal*), one of its members (*forest /tree*) or a substance made of (*pane/glass*).
- **Holonymy**: Relation binding a $concept_{-1}$ to a $concept_{-2}$ of which it is one of the parts. It is the opposite of the meronymy relation.
- **Implication**: Relation binding a $concept_{-1}$ to a $concept_{-2}$ which results from it (*to walk /take a step*).
- **Causality**: Relation binding a $concept_{-1}$ to its purpose (to kill /to die).

[1] https://wordnet.princeton.edu/wordnet/download/current-version/#win

- **Value**: Relation binding a $concept_{-1}$ (adjective) which is a possible state for a $concept_{-2}$ (*poor /financial condition*).
- **Has the value**: Relation binding a $concept_{-1}$ to its possible values (adjectives)(*size /large*). It is the opposite of relation value.
- **See also**: Relation between concepts having a certain affinity (*cold /frozen*).
- **Similar to**: Certain adjectival concepts which meaning is close are gathered. A synset is then designated as being central to the regrouping. The relation 'Similar to' binds a peripheral synset with the central synset (*moist /wet*).
- **Derived from**: Indicate a morphological derivation between the target concept (adjective) and the concept origin (*coldly /cold*).

We build TempoWordNets by enriching all the synset of WordNet 3.0 with its intrinsic temporal dimensions. In particular, all the synsets of WordNet 3.0 are automatically time-tagged with four dimensions: *atemporal*, *past*, *present,* and *future*.

We have adopted different strategies to build TempoWordNet. We will start by detailing the two-step strategy, which embodies most of the relevant concepts and then straightforwardly define the one-step process. It is pertinent to mention that these two strategies are inspired by the ideas of [Esuli and Sebastiani, 2005] and [Esuli and Sebastiani, 2006]. Subsequently, we will discuss two other strategies i.e. (i) the probabilistic expansion and (ii) the hybrid (probabilistic combined with semantic) expansion followed to build TempoWordNet.

3.3.1 Two-Step Classification

The overall idea of the two-steps strategy can be described as follows. First, a three-class temporal classifier is built over a set of manually selected seed synsets defined by their corresponding glosses. The underlying idea is that temporal synsets should embody temporality in their definition in a similar way. The classification process is iterated based on the repetitive semantic expansion of the initial seeds lists until

cross-validation accuracy drops. By semantic expansion, we mean that different lexico-semantic relations are used to encounter temporality in WordNet. This first step results in a *past*, *present* and *future* classifier and an expanded list of temporal synset candidates.

A second temporal classifier is then learned to time-tag synsets as *atemporal* or *temporal*. This process is obtained by taking the final list of expanded seed synsets from the previous learning problem and randomly choosing a balanced number *atemporal* synsets. A 10-fold cross-validation is then used to learn the model.

TempoWordNet is finally obtained by first classifying all WordNet synsets as *atemporal* or *temporal* with the second classifier and then the resulting temporal synsets are tagged as *past*, *present* and *future* by the first classifier.

Past, Present, Future Classification

The first step to build TempoWordNet is based on a classification model, which aims to distinguish between *past*, *present* and *future* synsets. This first step is defined in Algorithm 1 and all subtasks are explained as follows:

Algorithm 1 Past, present, future classification.

```
Selection of the initial seeds lists
repeat
  Expansion of the seeds lists
  Learning the model Past, Present, Future
  Measure Accuracy by 10-fold cross-validation
until accuracy drops
```

Initial Seeds Lists Selection: In SentiWordNet, [Esuli and Sebastiani, 2005] starts by selecting words that are relevant to express positive or negative opinions. Similarly, we need to select seeds used as good paradigms for *past*, *present* and *future* categories. For example, words like "yesterday", "previously", "remember" are good paradigmatic words for the *past* category, "current", "existing", "presently" for *present* and "prophecy", "predict", "tomorrow" for *future*. The selection of the

initial set of seed synsets is a crucial step in the process as their properties must be preserved along the expansion process. Therefore, any imperfect initial set choice will have huge consequences.

In order to catch the most relevant synsets for each time category, a first selection was made by several individuals through intensive and freewheeling group discussion. Every participant was encouraged to think aloud and to suggest as many words as possible. We preferred to use this process as choosing all words from the WordNet time subtree would have resulted in a biased sample as almost all synsets are nouns. Indeed, we wanted to make sure that each grammatical category existing in WordNet (i.e. Noun, Adjective, Adverb and Verb) would be present in the sets of seeds for *past*, *present* and *future* categories.

As each synset in WordNet contains one or more words, the synsets expressing the temporal connotations listed by the individuals were selected.

Finally, we performed an inter-annotator agreement process over the three seeds lists with four different annotators who were presented with the synsets and their respective glosses. The results of the multi-rater agreement evaluation are presented in Table 3.1. In particular, we processed the free-marginal multi-rater kappa values [Randolph, 2005] as Fleiss' popular multi-rater kappa [Fleiss, 1971] is known to be influenced by prevalence and bias, which can lead to the paradox of high agreement but low kappa. Overall figures assess adequate agreement.

Table 3.1 Inter-annotator agreement.

Metric	Past	Present	Future
% of overall agreement	0.85	0.83	0.90
Free-marginal κ	0.70	0.66	0.80

The initial lists of *past*, *present* and *future* seeds are given in Table 3.2.

Expansion Process: The guiding idea behind the expansion process is that the temporal properties of the initial hand-crafted seeds lists should be preserved as we

Table 3.2 List of 30 initial *temporal* seeds equally distributed over *past*, *present* and *future*.

Words	Sense	Category	Class
past	1,2	n.	past
past	1,2	adj.	past
yesterday	1,2	n.	past
yesterday	1,2	adv.	past
commemorate	2	v.	past
previously	1	adv.	past
present	1	n.	present
present	1,2	adj.	present
now	1	n.	present
now	3	adv.	present
nowadays	1	adv.	present
today	1	n.	present
ongoing	1	adj.	present
existing	1	adj.	present
current	1	adj.	present
future	1	n.	future
future	1,2	adj.	future
tomorrow	1,2	n.	future
tomorrow	1	adv.	future
predict	1	v.	future
expected	1	adj.	future
prophesy	1	v.	future
aforethought	1	adj.	future

strategically travel through WordNet. Depending on the morpho-syntactic class of an initial temporal synset, choosing an appropriate set of conceptual relations may allow to expand the notion of time in WordNet. Initially, we have tried several conceptual relations available in WordNet to expand initial seeds lists. We have also exploited combination of relations in order to expand the seeds list. However, the results were not encouraging. Finally, we decided to exploit the following conceptual relations for the expansion of initial hand crafted seeds lists depending on the morpho-syntactic class of each seed. Following relations are used for this purpose:

- **synonymy:** for each morpho-syntactic class (e.g. "past" vs. "yesteryear" for noun),
- **hyponymy:** for nouns[2] (e.g. "future" vs. "tomorrow"),
- **troponymy:** for verbs (e.g. "will" vs. "plan"),
- **related nouns:** for adjectives (e.g. "future" vs. "approaching"),
- **root adjectives:** for adverbs (e.g. "recently" vs. "recent").

Classification: Finally, a semi-supervised learning strategy is used to learn the temporal (*past*, *present* and *future*) classifier. At each semantic expansion step (or iteration), a three-class text classifier is trained over the glosses and lemmas of each synset contained in the seeds lists. After each iteration, the accuracy of the learned model is measured through a 10-fold cross-validation process. The expansion process continues until the classifier accuracy steadily drops.

Results: In our experiments, we used the initial seeds lists containing 30 synsets and then performed the semi-supervised learning process using different classifiers and representations. As for classifiers, we used Support Vector Machines (SVM), Multinomial Naïve Bayes models (MNB) and Decision Trees (C4.5) from the Weka plateform[3] and performed all the experiments with default parameters set by Weka. As for the representation space, each synset was represented by its gloss encoded as a vector of word unigrams weighted by their frequency in the gloss. Stop words removal has been performed using the Weka database.

Overall results are presented in Table 5.1 and show that the "optimal" expansion is obtained after three iterations using SVM. In the cases of MNB and C4.5, accuracy immediately drops as the introduction of possible noisy synsets is hard to handle for such simple models considering the small size of training data.

[2]Following the hypernymy relation leads to the classical semantic shift problem.
[3]http://www.cs.waikato.ac.nz/ml/weka/ [Last access: 09/09/2015].

Table 3.3 SVM, naïve bayes and decision trees accuracy results for Past, Present, Future classification at each iteration step.

Steps		1	2	3	4	5
SVM	Precision	81.7	84.4	86.1	86.0	85.4
	Recall	79.9	82.6	83.9	83.5	83.2
	F_1-measure	79.8	82.1	**83.5**	83.1	82.9
MNB	Precision	83.8	76.9	78.2	77.4	78.1
	Recall	82.7	76.7	77.5	76.3	77.0
	F_1-measure	**83.2**	76.8	77.8	76.8	77.5
C4.5	Precision	76.3	73.5	71.1	72.4	73.5
	Recall	70.8	63.5	63.5	63.8	62.5
	F_1-measure	**74.4**	68.1	67.1	68.4	67.6

Finally, we end up with a list of 632 temporal synsets distributed as follows: 210 synsets marked as *past*, 291 as *present* and 131 as *future*. In Table 3.4, we provide the top 10 synsets and the bottom 10 synsets classified as temporal by the SVM at iteration 3. Likewise the distribution of the number of extracted synsets, the distribution of morpho-syntactic categories depends on the temporal class. For instance, *future* is mainly referred to by nouns, while *present* evidences a high number of action verbs and *past* is represented by ancient animals or adjectives and adverbs.

These results were expected. However, when digging up results, the temporal issue of synsets is sometimes difficult to guess. In fact, it is important to remember that classification is made over glosses. As such, the temporal values of concepts are given by their definition. For instance, "here", which is classified as a *present* adjective has an unclear temporal connotation. However, its gloss "*being here now*" clearly refers to a *present* situation.

Also, within the expansion process, noisy synsets may be introduced and complicated the learning process. For instance, "augur", which is automatically defined as a noun with a *future* connotation is incorrectly classified. Indeed, its gloss "*a religious official who interpreted omens to guide public policy*" does not embody any

future issue. This situation is discussed later but has mainly to deal with the fact that the temporal connotation is not always present for a same denotation.

Atemporal vs. Temporal Classification

Once the *past*, *present* and *future* classifier has been learned, we end up with a list of 632 temporal synsets, which "abusively" embody the notion of time in WordNet. Indeed, there are more temporal categories than just *past*, *present* and *future* as there are more than positive and negative classes in the expression of sentiments. Although, as a first step towards building the first temporal ontology so far[4], we found wise to refer to the common sense connotations of time.

Continuing with the analogy proposed by [Esuli and Sebastiani, 2005] where any word is objective if it is not negative or positive, any concept, which is not associated to the notions of *past*, *present* or *future* is called *atemporal*. So, in order to learn the second classifier, we randomly chose a set 632 *atemporal* synset candidates within WordNet each one being outside the time subtree of WordNet and the list of pre-computed 632 temporal synsets. We performed a manual cross-annotation process to ensure the atemporality of the candidates. For that purpose, we randomly selected a subset of 10 synsets and asked four annotators to decide upon their atemporality. The results of the free-marginal multi-rater kappa evidence a substantial agreement with 0.73.

So, based on the set of 632 temporal synsets and 632 atemporal ones, a SVM was learned for a two-class problem reaching 85.6% accuracy over a 10-fold cross-validation process. Similarly to the first classification task, we used a linear kernel and represented each synset by its gloss based on the vector space model with each word feature being represented by its frequency.

Finally, TempoWordNet is obtained by a two-step process: (1) all synsets in WordNet are classified as *temporal* or *atemporal* based on the classifier mentioned in

[4]As far as we know.

subsection 3.3.1 and (2) each temporal synset is associated to its temporal values (*past*, *present* and *future* summing up to one) using the classifier built in subsection 3.3.1. Examples of time-tagged synsets with their numerical scores are presented in Table 3.5.

3.3.2 One-Step Classification

One direct comparison has been experimented with a one-step classification strategy. Instead of expanding temporal synsets in a first step to finally execute two-stage classification, propagation can be executed in a single step.

So, we propose to expand both *temporal* and *atemporal* synsets at the same time and directly produce a four-class temporal classifier: *past*, *present*, *future* and *atemporal*. For that purpose, we presented a set of 30 *atemporal* synsets to four annotators who agreed with a free-marginal multi-rater kappa value over 0.8 indicating almost perfect agreement. The list of *atemporal* synsets is given in Table 3.6.

The same semi-supervised learning strategy is used to learn the four-class (*past*, *present*, *future* and *atemporal*) classifier. At each iteration, the classifier is trained over the glosses and lemmas of each synset contained in the seeds lists. After each iteration, accuracy is measured through a 10-fold cross-validation process and the expansion process stops when accuracy drops. Results are presented in Table 3.7 for the same experimental setups as for the two-step strategy.

Unlike the two-step strategy, the one-step process shows incapacity to solve the temporality issue. Indeed, introducing the *atemporal* synsets in the propagation process since the first iteration evidences different problems: (1) atemporality is difficult to define unless when opposed to temporality as it embodies many denotations and no connotation and (2) temporality only spreads over a small proportion of WordNet while atemporality covers most of WordNet, and as such, as iterations grow, the set of atemporal candidate synsets gets predominant (i.e. more unbalanced datasets are obtained after each iteration).

It is evident from the previous sections that *Two-Steps Classification* approach is performing better than *One-Step Classification* for solving temporal issues. However, it has some limitations. As the expansion process is semantically driven, the temporal connotation is highly dependent on the initial seeds lists and as a consequence may not spread over a wide range of concepts in WordNet. Therefore, we propose two different strategies of expansion: (1) the probabilistic expansion and (2) the hybrid (probabilistic approach combined with semantic) expansion.

3.3.3 Probabilistic Expansion

Due to small amount of hand labeled training data [5], we adopted a dynamic learning framework which requires a small amount of labeled data in the beginning, then incrementally discovers confidently classified unlabeled data and incorporates them into the training set to improve learning performance as well as to increase lexical coverage. This approach has great potential to reduce the time required to build a large set of manually labeled training data. It is also beneficial considering that not all labeled data have the same level of effectiveness in improving a classifier. In our experiment we consider the confidence rated classifiers that can predict a probability distribution over the labels for an example since the probability distribution enables us to determine the "confidence" of classification for the example.

We first learn a *temporal* vs. *atemporal* classifier based on the initial hand-crafted set of seeds. In particular, the seeds defined as *past*, *present* and *future* are markers of temporality, while the list of *atemporal* synsets is the obvious counterpart. Based on this list of *temporal* and *atemporal* synsets, a 10-fold cross validation process is performed to learn the *temporal* vs. *atemporal* model, which is used to time-tag the whole WordNet. The synsets (or glosses) with highest *temporal* and *atemporal* values in WordNet are then used for the expansion process of the seeds lists. The process is iteratively performed and stops when accuracy drops. The overall process is presented in Figure 3.1.

[5] 30 temporal synsets and 30 atemporal synsets

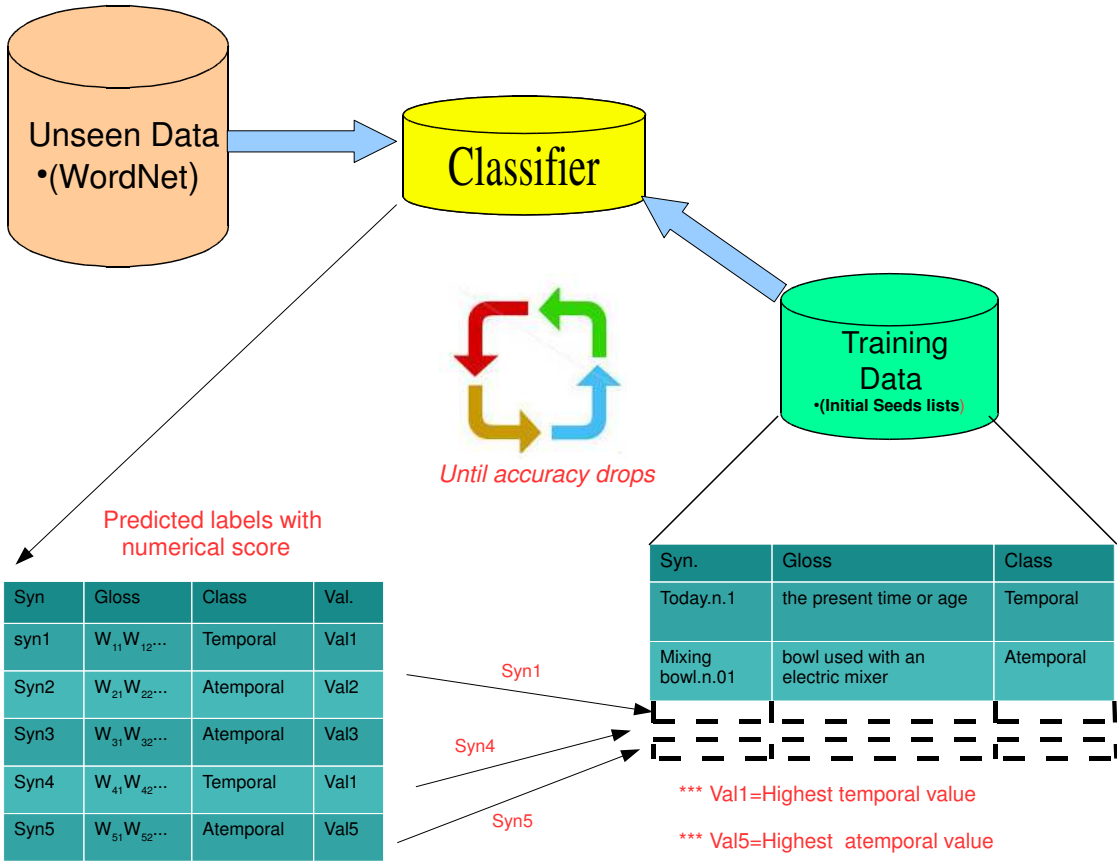

Figure 3.1: Probabilistic Expansion Strategy

After building the *temporal* vs. *atemporal* classifier, WordNet is divided into two subsets: *temporal* synsets and *atemporal* ones. In order to fine tune the *temporal* part of WordNet, we learn a three-class classifier (i.e. *past*, *present* and *future*) based on the initial *past*, *present* and *future* seeds lists and the probabilistic expansion exclusively within the temporal part of WordNet. In particular, temporal synsets are classified as *past*, *present* or *future* and used for the expansion process. Like the previous strategy, 10-fold cross validation process is iteratively performed until accuracy drops.

The results of the probabilistic expansion are presented in Table 3.8 and Table 3.9, when the expansion is based on the maximum probability value[6]. Examples from the final resource obtained by the probabilistic expansion strategy are presented in Table 3.5.

Note that in our experiment, Support Vector Machines (SVM) with a linear kernel[7] over the vector space model representation of the synsets (i.e. each synset is represented by its gloss encoded as a vector of unigrams weighted by their frequency) have been used to classify all the synsets of WordNet. The results show that in both cases the expansion process stops at iteration 2.

3.3.4 Hybrid Expansion

Choosing synsets from WordNet with highest probability assigned by a classifier learned on the glosses of initial seeds lists can lead to the well-known semantic shift problem. So, the idea of the hybrid expansion is to control the expansion process so that the most probable time-sensitive synsets are also chosen based on their semantic distance with the expanded seed synsets at the previous iteration. The process is straightforward when compared to the probabilistic expansion. An outline of the overall hybrid strategy is depicted at Figure 3.2.

First, a two-class (*temporal* vs. *atemporal*) text classifier is trained based on the glosses of each synsets contained in the initial seed lists to classify all the synsets of WordNet. Thereafter, WordNet synsets with highest probability are selected as candidates for expansion. From these candidates, only the ones that present the maximum semantic similarity to the previous seeds lists are chosen for expansion. Note that the semantic similarity is calculated between the candidate synset and all synsets in the previous expanded seeds lists. Once candidates for expansion have been chosen, a 10-fold cross validation process is iteratively performed until accuracy becomes steady.

[6]That means that all the synsets getting the highest value produced by the classifier are used to expand the initial seeds lists.

[7]We used the Weka implementation SMO with default parameters.

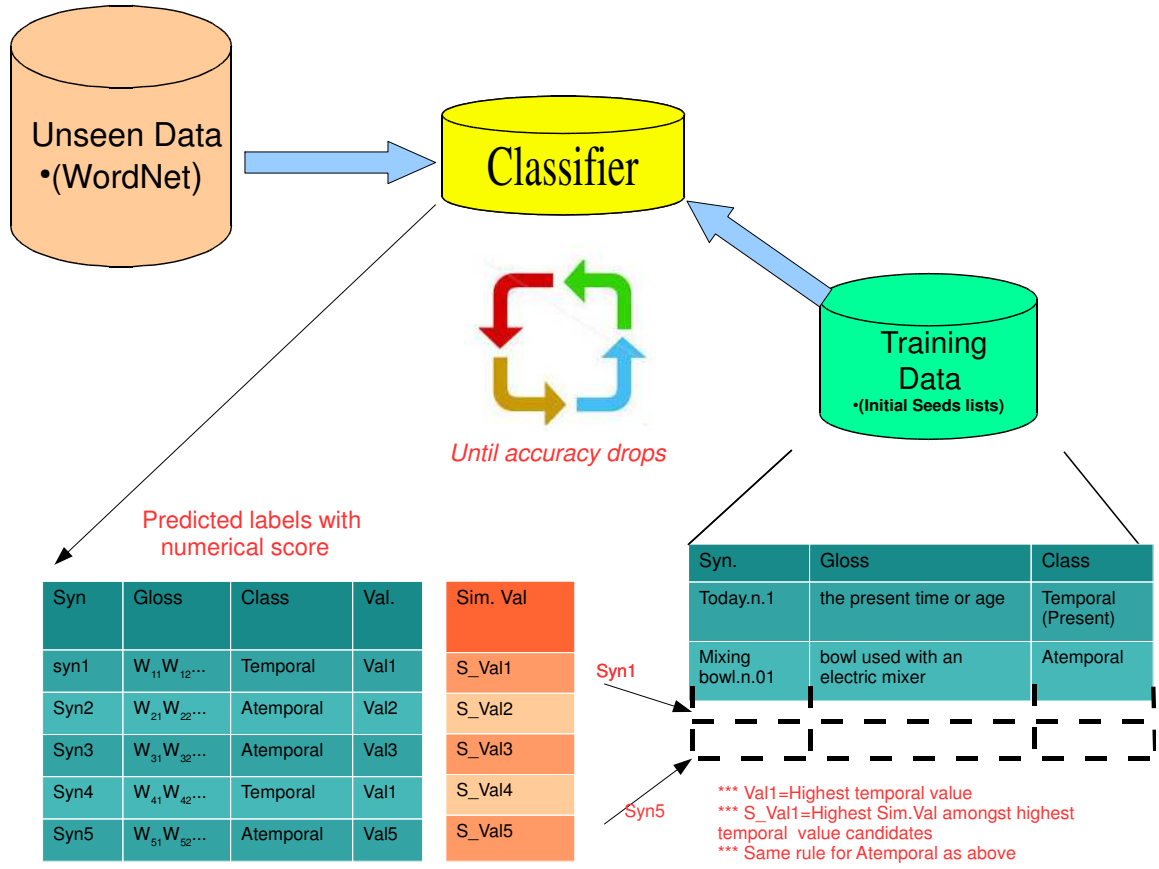

Figure 3.2: Hybrid Expansion Strategy

Second, a three-class (*past*, *present* and *future*) classifier is learned over the *temporal* part of WordNet with the hybrid expansion process in the same exact manner as explained for the previous probabilistic expansion. Results for the expansion process are presented in the Table 3.10 and Table 3.11 for the same experimental setups as for the probabilistic expansion and using the Leacock and Chodorow (lch) semantic similarity measure [Leacock et al., 1998]. Entries from the lexical resource obtained by the hybrid expansion process are presented in Table 3.5.

3.4 Evaluation

In this section we perform two experiments to asses the quality of the different versions of TempoWordNets: (i) Manual Evaluation (ii) Automatic Evaluation. The details of each strategies are as follows.

3.4.1 Manual Evaluation

In order to intrinsically evaluate the different versions of TempoWordNets, we first performed an inter-annotation process over samples of 50 automatically time-tagged WordNet synsets. In particular, three different annotators were presented with *temporal* synsets and their respective glosses, and had to decide upon their correct classification (*temporal* vs. *atemporal*). The results of the multi-rater agreement evaluation are presented in Table 3.12. In particular, we processed the free-marginal multirater kappa values [Randolph, 2005] and the fixed-marginal multirater kappa [Siegel and Castellan, 1988] as no bias is present in the data. Overall figures assess moderate agreement for the three TempoWordNets: TWnL for the lexico-semantic expansion, TWnP for the probabilistic expansion and TWnH for the hybrid expansion.

These results evidence the difficulty of the task for humans as they do not agree on a great deal of decisions. This is particularly due to the fact that the temporal dimensions of synsets are judged upon their glosses and not directly on their inherent concept. For example, "dinosaur" can be classified as *temporal* or *atemporal* as its gloss *any of numerous extinct terrestrial reptiles of the Mesozoic era* allows both interpretations.

So, we performed a new experiment based on those examples where human annotator agreement was 100%. From this dataset, we performed an inter-annotator agreement process with four annotators (three human annotators plus the classifier). The underlying idea is to understand to what extent the built TempoWordNets complies with the "easy" cases. Results are illustrated in Table 3.13 and clearly show the

enhanced intrinsic quality of the hybrid expansion strategy with an almost adequate agreement for the free-marginal κ.

3.4.2 Automatic Evaluation

In order to evaluate TempoWordNets, we propose to evidence their usefulness based on an external task: sentence temporal classification. The underlying idea is that a temporal knowledge base can help to classify sentences into three different categories: *past*, *present* and *future*.

For that purpose, we automatically selected a set of *past*, *present* and *future* sentences from the well-known SemEval-2007 corpus developed for task 15 [Verhagen et al., 2007]. This corpus is a version of TimeBank containing approximatively 2500 sentences with TimeML annotations. So, all sentences exclusively containing *past* (resp. *present*) expressions were marked as *past* (resp. *present*). As for *future*, all sentences containing *future* expressions combined or not with *present* timexes were tagged as *future*. The final corpus consists of 1455 sentences distributed as follows: 724 for *past*, 385 for *present* and 346 for *future*. Some examples are given as follows:

1. *In New York Stock Exchange composite trading yesterday, Oneida's shares closed at $18.375 a share, unchanged* (**Past**),
2. *Currently, Avon, based in Santa Monica, Calif., has 3.3 million common shares outstanding* (**Present**),
3. *A TOP-LEVEL investigation into Mark Thatcher's alleged arms deals with Iraq is to be launched by Swiss government officials in the New Year* (**Future**).

Different sentence representations have been used. First, we proposed to represent each sentence with the classical vector space model using the tf.idf weighting scheme for unigrams without stop-words removal (Uni.+SW). Then, we proposed a semantic vector space representation where each sentence is augmented with the synonyms

of any temporal word contained in it. In particular, we proposed that the words were matched directly from the WordNet time subtree (Uni.+SW+Wn) or from TempoWordNet (Uni.+SW+TWnL, Uni.+SW+TWnP and Uni.+SW+TWnH) and weighted with tf.idf. The results of our experiments are reported in Table 3.14. The results evidence that the WordNet time subtree does not embody enough time-related information and the process of automatically time-tagging WordNet can improve the task of sentence temporal classification, especially with the probabilistic or the hybrid expansion.

In order to better understand the process of temporal classification of English sentences, we propose to compare the unigram representations with and without stop words. Results are shown in Table 3.15.

Stop words indeed play an important role for sentence classification for the english language. In the list of stop words there are auxiliary verbs such as "will" or "did", which are evident clues for sentence temporal classification. The improvement of TempoWordNet on this small dataset is therefore residual reaching 2.2% increased performance. However, its importance may not be neglected as complex temporal classification tasks are not likely to depend on auxiliary verbs.

3.5 Summary

We proposed the first steps towards automatic construction of a temporal ontology. In particular, we build TempoWordNet, a lexical resource in which each WordNet synset is associated with four numerical scores. These scores reflect temporal orientation of the terms contained in the synset in terms of *past*, *present*, *future*, and *atemporal*.

The proposed approach to construct TempoWordNet relies on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification. First, synsets are classified as *atemporal* or *temporal* and then, all temporal synsets are associated to *past*, *present* and *future* scores.

In order to evaluate the usefulness of TempoWordNets, different strategies are adopted. A sample of randomly selected TempoWordNet entries are manually evaluated. For automatic evaluation, we propose to evidence its usefulness based on sentence temporal classification. The underlying idea is that a temporal knowledge base can help to classify sentences into three different categories: *past*, *present* and *future*. The experimentation shows that the resource can greatly improve the sentence temporal classification task. We also examined the role of stop words in English sentence temporal classification.

We deeply believe that TempoWordNet can be an important resource for time related applications both in NLP and IR. As a consequence, we provide free access to this resource as well as all developing materials at `https://tempowordnet.greyc.fr/`. A screenshot of the website is given in Figure 3.3.

Figure 3.3: Screenshot of TempoWordNet Website

Table 3.4 List of automatically retrieved temporal synsets.

Past			
Top 10		Bottom 10	
Word (Sense)	Cat.	Word (Sense)	Cat.
by (1)	adv.	iguanodon (1)	n.
recently (1)	adv.	ground-shaker (1)	n.
in the first place (1)	adv.	diplodocus (1)	n.
remember (3)	v.	saurischian (1)	n.
old (1)	n.	argentinosaur (1)	n.
old (1)	adj.	ornitischian (1)	n.
old (2)	adj.	titanosaur (1)	n.
old (6)	adj.	mellowing (1)	n.
erstwhile (1)	adj.	appearance (4)	n.
honest to god (1)	adj.	psychosexuality (1)	n.
Present			
Top 10		Bottom 10	
Word (Sense)	Cat.	Word (Sense)	Cat.
immediately (1)	adv.	overstay (1)	v.
now (3)	adv.	visit (7)	v.
presently (2)	adv.	run (30)	v.
present (3)	n.	drag on (1)	v.
immediate (3)	adj.	wear (6)	v.
instant (2)	adj.	crawl (3)	v.
attendant (1)	adj.	bond (3)	v.
ever-present (1)	adj.	ramp (5)	v.
here (1)	adj.	stand back (2)	v.
omnipresent (1)	adj.	line up (3)	v.
Future			
Top 10		Bottom 10	
Word (Sense)	Cat.	Word (Sense)	Cat.
prophecy (1)	n.	example (4)	n.
prefiguration (2)	n.	referral (1)	n.
prognosis (1)	n.	palmist (1)	n.
prophecy (2)	n.	sibyl (1)	n.
meteorology (1)	n.	prophetess (1)	n.
fortunetelling (1)	n.	augur (1)	n.
extropy (1)	n.	sibyl (2)	n.
horoscope (1)	n.	onomancy (1)	n.
guess (2)	n.	arithmancy (1)	n.
credit rating (1)	n.	lithomancy (1)	n.

Table 3.5 Examples of time-tagged synsets with their numerical scores

Numerical Scores		Past	Present	Future	Atemporal
TWnL	bygone.n.01	**0.997002**	0.001998	0.000000	0.000000
	time_being.n.01	0.000000	**0.992012**	0.001988	0.006000
	prefiguration.n.02	0.000000	0.001988	**0.992012**	0.006000
	shy.s.03	0.000000	0.009980	0.000020	**0.990000**
TWnP	bygone.n.01	0.017000	0.087000	**0.896000**	0.000000
	time_being.n.01	0.002000	**0.998000**	0.000000	0.000000
	prefiguration.n.02	0.009000	0.088000	**0.903000**	0.000000
	shy.s.03	0.000000	0.000000	0.000000	**1.000000**
TWnH	bygone.n.01	**0.990000**	0.000000	0.010000	0.000000
	time_being.n.01	0.000000	**0.998000**	0.002000	0.000000
	prefiguration.n.02	0.002000	0.000000	**0.998000**	0.000000
	shy.s.03	0.000000	0.000000	0.000000	**1.000000**

Table 3.6 List of initial *atemporal* seeds.

Words	Sense	Category	Class
mixing bowl	1	n.	atemporal
freshen	2	v.	atemporal
carnation	2	n.	atemporal
chadian	1	adj.	atemporal
wren warbler	1	n.	atemporal
brainsick	1	adj.	atemporal
estriol	1	n.	atemporal
theology	2	n.	atemporal
unexpectedly	1	adv.	atemporal
jabber	1	n.	atemporal
human waste	1	n.	atemporal
cruciferous	1	adj.	atemporal
pet sitter	1	n.	atemporal
trombicula	1	n.	atemporal
drum	1	v.	atemporal
dateline	1	n.	atemporal
shot	11	n.	atemporal
okinawa	1	adv.	atemporal
chatter	1	v.	atemporal
polecat	2	n.	atemporal
foster home	1	n.	atemporal
lymph node	1	n.	atemporal
arabian sea	1	n.	atemporal
semanticist	1	n.	atemporal
strauss	3	n.	atemporal
doric order	1	n.	atemporal
reptantia	1	n.	atemporal
belt	2	v.	atemporal
half dollar	1	n.	atemporal
staggered board of directors	1	n.	atemporal

Table 3.7 SVM, naïve bayes and decision trees accuracy results for Past, Present, Future, Atemporal classification at each iteration step.

Steps		1	2	3	4	5
SVM	Precision	81.3	68.0	71.0	71.3	71.7
	Recall	80.3	63.0	66.8	67.8	68.6
	F_1-measure	**80.8**	65.4	68.8	69.5	72.8
BNB	Precision	75.2	67.0	67.1	67.8	77.4
	Recall	74.3	64.6	63.6	64.0	76.3
	F_1-measure	74.7	65.8	65.3	65.8	**76.8**
C4.5	Precision	73.3	68.6	59.2	61.4	72.4
	Recall	70.6	52.4	48.6	48.0	63.8
	F_1-measure	**71.9**	59.4	53.3	68.4	69.9

Table 3.8 Cross validation for *temporal* vs. *atemporal* at each iteration. Probabilistic Expansion.

Steps	1	2	3
Precision	87.3	**100**	100
Recall	86.7	**100**	100
F_1-measure	86.9	**100**	100

Table 3.9 Cross validation for *past*, *present* and *future* at each iteration. Probabilistic Expansion.

Steps	1	2	3
Precision	80.0	**99.7**	99.6
Recall	80.1	**99.7**	99.6
F_1-measure	80.0	**99.7**	99.6

Table 3.10 Cross validation for *temporal* vs. *atemporal* at each iteration. Hybrid Expansion.

Steps	1	2	...	25	26	27
Precision	87.3	94.1	...	96.0	**97.2**	96.6
Recall	86.7	93.2	...	95.5	**97.0**	96.3
F_1-measure	86.9	93.6	...	95.7	**97.1**	96.4

Table 3.11 Cross validation for *past*, *present* and *future* at each iteration. Hybrid Expansion.

Steps	1	2	...	15	16	17
Precision	80.0	75.7	...	95.7	**96.4**	95.6
Recall	80.1	74.3	...	95.1	**96.0**	95.0
F_1-measure	80.0	74.9	...	95.4	**96.2**	95.3

Table 3.12 Inter-annotator agreement.

Metric	TWnL	TWnP	TWnH
Fixed-marginal κ	0.5073	0.5199	0.4197
Free-marginal κ	0.5199	0.5199	0.4399

Table 3.13 Inter-annotation for "easy" cases.

Metric	TWnL	TWnP	TWnH
Fixed-marginal κ	0.4133	0.4767	0.5655
Free-marginal κ	0.4242	0.5161	0.6896

Table 3.14 Evaluation results for sentence classification with different TempoWordNets. Balanced corpus with stop words: 346 sentences for *past*, 346 sentences for *present* and 346 sentences for *future*.

Representation	Uni. +SW	Uni.+SW +Wn	Uni.+SW +TWnL	Uni.+SW +TWnP	Uni.+SW +TWnH
Precision	85.8	85.6	87.8	**89.8**	89.5
Recall	85.7	85.3	87.8	**89.5**	89.4
F_1-measure	85.8	85.4	87.8	**89.6**	89.4

Table 3.15 Evaluation results for sentence classification without stop words. Balanced corpus: 346 sentences for *past*, 346 sentences for *present* and 346 sentences for *future*

	Uni.	Uni.+TWnL
Precision	64.2	78.3
Recall	64.3	77.8
F_1-measure	64.2	**78.0**

Chapter 4

Application of TempoWordNet

In the previous chapters we studied how time is expressed in natural language text and introduced a temporal ontology to understand better the language of time. We built TempoWordNet where each synset is associated to its intrinsic temporal dimensions: *atemporal, past, present, and future*. We also showed how to improve sentence temporal classification tasks by incorporating temporal knowledge from the resource.

Now that the resource is in place, to complement our research we can take a look at how a search application can benefit from this resource. In this chapter, we examine temporal intent behind a user's search query. In particular, we present our framework to tackle the problem where TempoWordNet plays crucial role.

This chapter is organized as follows. The following section presents a general introduction to underlying temporal intent behind user's search query. Their categorizations are also covered in the same section. Afterwards, Section 4.2 presents closely related works on query classification. Section 4.3 introduces the gold standard data sets and resources used for our experiments. The overall approach to classify temporal intent of queries are covered in Section 4.4. Section 4.5 presents the experimental set up and results. Finally, importance of different features used in the learning process are examined in Section 4.6.

4.1 Overview of Temporal Query Intent

Web is a dynamic information source in which the number and content of pages change continuously over time. Web search queries are dynamic in nature and temporally sensitive. Temporally sensitive implies that the intent of a given user for information changes over time. Many queries may only be answered accurately if their underlying temporal orientations are correctly judged. So, recognizing the temporal intent behind users' queries is a crucial part towards improving the performance of information access system and also diversifying the retrieved results from a search engine. For instance, this can be useful to select specific temporal retrieval models [Li and Croft, 2003], temporally re-rank web results [Kanhabua and Nørvåg, 2012] or assess credible information [Schwarz and Morris, 2011].

Depending on the temporal intent, web search queries can be classified into two main categories: *explicit* temporal query and *implicit* one. Explicit temporal queries are the ones that contains exclusive temporal expressions inside query keywords. Some examples are *Olympic Games 2012, 2018 FIFA World Cup etc.* Despite an apparent timeless nature, implicit temporal queries embody inherent temporal evidence. They consist of a set of keywords implicitly related to a particular time interval, which is not explicitly specified by the user. Some examples are *Nepal earthquake, stock price of Apple, long term weather forecast for Paris etc.*

Many web search queries have implicit or explicit intents related with them. According to the survey performed over AOL query dataset [Metzler et al., 2009], 1.5 % of queries have explicit temporal intents. While according to [Metzler et al., 2009], the rate of implicit temporal queries is more than 7% . Considering the importance of web search in today's life, the rate of queries with implicit or explicit temporal intents amount to huge number of searches everyday. These queries may only be answered accurately if their underlying temporal orientations are correctly judged. A user who enters the query "Apple-IOS8" may wish to find the official web page for the mobile operating system, reviews about the operating system, or the release date of IOS8. However, it may be a challenging task to accurately determine which of

these implicit intents user actually meant since the query has only two keywords. We focus on queries that are particularly of temporal nature rather than concentrating on the problem of automatically determining user intent for generic one.

Similar to [Joho et al., 2014a], we considered four temporal classes namely past, recency, future, and atemporal depending on the implicit or explicit temporal aspect of query strings. The plausible definitions of these four temporal classes according to NTCIR [Joho et al., 2014a] are as follows:

- **Past:** Query related to events which are gone by in time. Search results are expected not to change much along with time passage (e.g. "*Who Was Martin Luther*", "*Yuri Gagarin Cause of Death*" etc.).
- **Present:** Query related to recent events, and returns up to date search results (usually this type of query refers to events that happened in very near past or at present time, unlike the âĂIJpastâĂİ query that tends to refer to events in relatively distant past). The information contained in search results usually changes quickly along with the time passage (e.g. "*time in Paris*", "*did Barcelona Win Today*" etc.).
- **Future:** Query about predicted or scheduled events, the search results of which should contain future-related information (e.g."*long term weather forecast*" , "*2018 FIFA World Cup schedule*" etc.).
- **Atemporal:** Query without any clear temporal intent (its returned search results are not expected to be related to time). Navigational queries are considered to be atemporal (e.g."*lose weight quickly*" ,"*New York Times*" etc.).

In this chapter, we present multi-objective optimization based ensemble learning paradigm to solve the problem of Temporal Query Intent Classification (TQIC). The task can be defined as follows.

Given a web search query q and its issuing date d, predict its temporal class $c \in \{past, recency, future, atemporal\}$.

4.2 Related Work

We now briefly discuss the related works that are closely related to the query classification task.

The most influential attempt at temporal classification of queries comes from [Jones and Diaz, 2007] who classify queries into three distinct classes: *atemporal*, *temporally unambiguous* and *temporally ambiguous*. In particular, they consider the distribution of retrieved documents over time and create meaningful features based on this distribution. Their classification is then done on this feature set. As an example, the time distribution of retrieved documents for the query "iraq war" has distinct peaks at the years 1991 and 2003 and is therefore classified as temporally ambiguous, while a query like "poaching" is classified as atemporal since its time distribution does not exhibit any peaks. A weakness of their approach is that only publication times of documents are considered. Often this publication time differs from the actual content time. Other ideas for implicit temporal queries have been developed by [Metzler et al., 2009]. By analyzing query logs, they investigate the automatic detection of implicitly year qualified queries, i.e. queries that refer to an event in a specific year without containing the year in the query string. Following the same motivation, [Campos et al., 2012a] proposed a solution based on content temporal analysis. In particular, they identify top relevant dates in web snippets with respect to a given implicit temporal query and temporal disambiguation is performed through a distributional metric called GTE.

Recently, the NTCIR Temporalia task [Joho et al., 2014b] pushed further this idea and propose to distinguish whether a given query is related to *past*, *recency*, *future* or *atemporal*. Within this context, the most performing system is based on a SVM semi-supervised learning algorithm [Yu et al., 2014] and uses the AOL 500K User Session Collection [Pass et al., 2006] as unlabeled data. Two other competitive systems [Shah et al., 2014, Hou et al., 2014] rely on ensemble learning (especially majority voting). Indeed, due to the small size of training data (only 100 queries distributed equally by class), classification results are weak if a single classifier is used

in a traditional supervised way. To overcome this situation, we follow the ensemble learning paradigm defined as a multi-objective optimization problem in a similar way as [Saha and Ekbal, 2013].

4.3 Learning Instances for TQIC

Although, there has recently been an increased attention in investigating temporal characteristics of queries, very few works exist that address user query's temporal intent. The Temporal Information Access [Joho et al., 2014b] is the first such challenge, which is organized to provide a common platform for designing and analyzing time-aware information access systems. It is hosted by the 11th NTCIR Workshop on Evaluation of Information Access Technologies (NTCIR-11)[1]. Organizers released one hundred (100) queries along with their respective temporal class and issuing time as training data (25 queries for each class). Three hundred (300) queries along with their issuing time were released as test data (75 queries for each class). We will call this data set NTCIR-TQIC. Examples of the form $< q, d, c >$ are given as follows.

q	d	c
who was martin luther	Jan 1, 2013 GMT+0	*past*
amazon deal of the day	Feb 28, 2013 GMT+0	*recency*
release date for ios7	Jan 1, 2013 GMT+0	*future*
number of neck muscles	Feb 28, 2013 GMT+0	*atemporal*

4.3.1 External Resources

The NTCIR-TQIC data set evidences two crucial limitations. First, the training set is small. Second, the amount of literal features is limited as queries are short (between 3 and 4 words). The first case is a classical learning problem and it is discussed in section 4.4. As for the second case, external resources were used to expand the amount of query information [Campos et al., 2012a].

[1] http://research.nii.ac.jp/ntcir/ntcir-11/

So, for each query, we first collected the top K web snippets[2] returned by the Bing search API[3]. The underlying idea is that web snippets are likely to evidence temporal information if the query has a temporal dimension. Some examples of the collected web snippet are presented in Table 4.1.

[2]For computational reasons, we set $K = 10$.

[3] `https://datamarket.azure.com/dataset/bing/search`

Table 4.1 Examples of urls and web snippets for given queries.

Queries	**Urls and Web Snippets**
michael douglas cancer	`http://www.cnn.com/2013/10/14/health/michael-douglas-tongue-cancer/index.htm.` Michael Douglas never had throat cancer, as he told the press in 2010. He actually had tongue cancer http://www.people.com/people/article/0,,20745023,00.html. The world was shocked when Michael Douglas announced he had stage four throat cancer in August 2010, but the Oscar winner now reveals that he..
price of samsung galaxy note	`http://mobiles.pricedekho.com/mobiles/samsung/samsung-galaxy-note-price-p4s8R.html.` Connectivity Offered. Samsung Galaxy Note offers a many connectivity options to the user, thus enabling them to get connected with other networks and devices within a ...http://www.mysmartprice.com/mobile/samsung-galaxy-note-msp1479 The best price of Samsung Galaxy Note in India is Rs. . The price has been sourced from 9 online stores in India as on 2014 15th June. The same price may be used to ...
i am a gummy bear	`http://www.youtube.com/watch?v=astISOttCQ0.` From the CD I Am Your Gummy Bear. Also from the DVD I Am A Gummy Bear Available on Amazon at: http://tinyurl.com/gummybeardvd ...http://www.youtube.com/watch?v=Z47EUaIFrdQ. My version of a log Gummybear video. Im no pro at mixing music, hope you like.Ive just finished making Mix 2 in which Ive removed the POP and fixed the ....
madden 2014 release date	`http://www.nflschedule2014.org/madden-release-date.html.` Madden 2015 Release Date The Madden NFL 15 release date is currently slated for August 30th 2014. Prelaunch information and developer leaks regarding Madden 15 will ...http://www.ign.com/articles/2014/04/28/madden-nfl-15-release-date-revealed. EA announced today that Madden NFL 15 is slated for release August 26 in North America and August 29 in Europe. The release date was revealed in the first ...

Then, for each query, we collected its most relevant year date along with its confidence value from the freely available web service GTE[4] proposed in [Campos et al., 2012a]. In particular, given a temporally implicit query, GTE extracts from web snippets query-relevant year dates based on distributional similarity. Some examples of extracted year dates and confidence values are given as follows[5].

q	Most confident Year	Confidence value
who was martin luther	1929	0.944
amazon deal of the day	2015	0.760
release date for ios7	2013	0.893
number of neck muscles	2014	0.708

4.3.2 Features Definition

The most important step in building a classifier is deciding what features of the input instances are relevant and how to represent them. Therefore, choosing discriminating and independent features are keys to any machine learning algorithm being successful in classification.

We identified eleven (11) independent features and used them in the learning process of our selected classifiers. These features are computed from the information extracted from three different sources and resources. All the considered features are listed in Table 4.2.

[4] `http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server/`
[5] Extraction was processed April, 16th 2015 for illustration.

Table 4.2 Overall features considered for temporal query intent classification.

Features	Description
D_b_Dates	Difference between dates.
C_o_Date	Confidence on date.
N_o_PaW	Number of Past words present in the query.
N_o_RW	Number of Recency words present in the query.
N_o_FW	Number of Future words present in the query.
N_o_PaS	Number of snippet classified as Past.
N_o_RS	Number of snippet classified as Recency.
N_o_FS	Number of snippet classified as Future.
N_o_AS	Number of snippet classified as Atemporal.
C_o_Q	Class of the query itself.
Q_S	Text of the query (unigrams).

Details of the different features are given as follows.

D_b_Dates: This feature aims to evaluate the time gap between the query and its issuing date. It is calculated as the difference between the year date explicitly mentioned in the query string q and the issue year date d_{year}. If there is no mention of a date inside q (timely implicit query), we consider the most confident year date obtained from GTE [Campos et al., 2012a]. If no date is returned by GTE, this feature is given a null value.

C_o_Date: This feature aims to evidence the confidence value over the time gap definition. It is set to 1 when there is explicit mention of a year date inside q string (maximum confidence). Otherwise, it is set to the returned confidence value of GTE [Campos et al., 2012a]. A 0 value is given if no date is returned by GTE.

N_o_PW, **N_o_RW** and **N_o_FW:** These features aim to capture the query timeliness based on its temporal content words. They respectively represent the number of words in q belonging to past, present and future categories in Tem-

poWordNet. In particular, we used TempoWordNet obtained by lexico-semantic expansion process.

N_o_PS, **N_o_RS**, **N_o_FS** and **N_o_AS:** This set of features aims to interpret the timeliness of the query q based on the temporality of its returned web snippets. The rationale is that if a query has a temporal dimension, web search results should evidence the same intent. So, for any q, these features are respectively the number of returned web snippets classified as past, recency, future and atemporal by the Sentence Temporal Classifier (STC) obtained at the time of building TempoWordNet.

C_o_Q: The aim of this feature is to define the intrinsic temporality of a query (as if it was a sentence). So, this feature takes the value returned by STC (i.e. past, recency, future or atemporal) when taking the query string q as input.

Q_S: The rationale of this feature is that specific (non-temporal) words may play an important role in temporal classification. As a consequence, each query string q is represented as its bag of unigrams where the presence of a word is associated to the value 1 and 0 when it is not present.

4.4 Learning Framework

An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or binary voting) to classify new examples [Dietterich, 2000]. In particular, ensemble learning is known to obtain highly accurate classifiers by combining less accurate ones thus allowing to overcome the training data size problem. Many methods for constructing ensembles have been developed in the literature [Dietterich, 2000]. For TQIC task, we propose to define ensemble learning as a multi-objective optimization (MOO) problem. Our motivations are twofold. First, [Saha and Ekbal, 2013] showed that MOO strategies evidence improved results when compared to single objective solutions and state-of-the-art baselines.

Second, MOO techniques propose a set of performing solutions rather than a single one. As TQIC can be thought as an intermediate module in some larger application (e.g. retrieval, ranking or visualization), offering different performing solutions can be a great asset to adapt to any kind of information access situation without loss of reliability.

4.4.1 MOO Problem Definition

Definition of Multi-Objective Optimization is stated as below:

Find the vector $\overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T$ of decision variables that optimize O objective functions simultaneously

$$\{OB_1(\overline{x}), OB_2(\overline{x}), \ldots, OB_O(\overline{x})\}$$

which also satisfy user-defined constraints, if any.

The concept of domination is an important aspect of MOO. In case of maximization of objectives, a solution $\overline{x_i}$ is said to dominate $\overline{x_j}$ if $\forall k \in 1, 2, \ldots, O,\ OB_k(\overline{x_i}) \geq OB_k(\overline{x_j})$ and $\exists k \in 1, 2, \ldots, O,$ such that $OB_k(\overline{x_i}) > OB_k(\overline{x_j})$. Among a set of solutions SOL, the non-dominated set of solutions $SOL^{'}$ are those which are not dominated by any member of the set SOL. The non-dominated set of the entire search space S is called the globally Pareto-optimal set or Pareto front. In general, a MOO algorithm outputs a set of solutions not dominated by any solution encountered by it.

These notions can be illustrated by considering an optimization problem with two objective functions — say, OB_1 and OB_2 — with six different solutions, as shown in Figure 4.1. Here target is to maximize both of the objective functions OB_1 and OB_2. In this example, solutions 3, 4 and 5 dominate solutions 1, 2 and 6. Solutions 3, 4 and 5 are non-dominating to each other. Because solution 3 is better than solution 4 with respect to objective function OB_1, but solution 4 is better than solution 3 with respect to OB_2. Similarly solution 4 is better than solution 5 with respect to OB_1

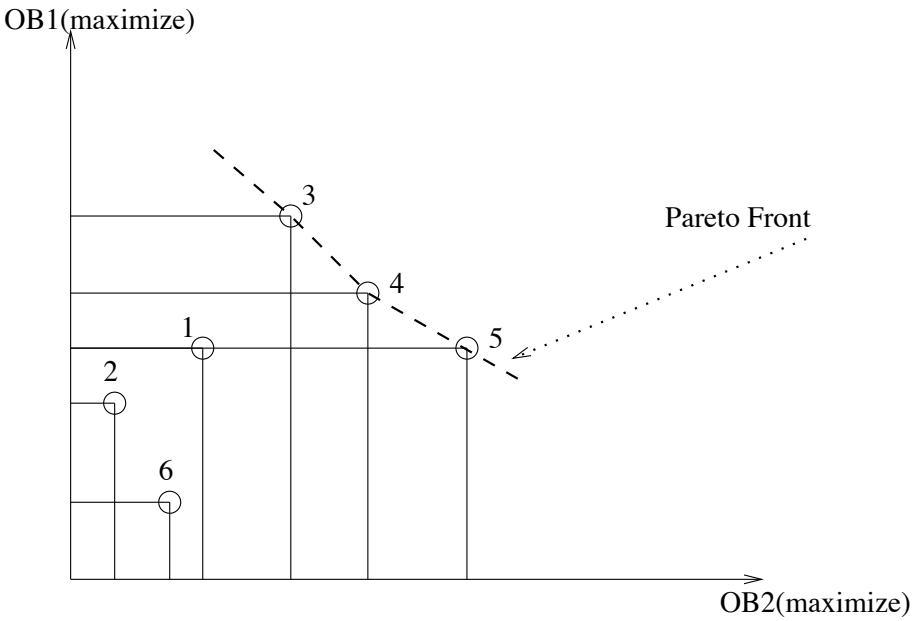

Figure 4.1: Example of dominance and non-dominance in MOO and Pareto-optimal-front

but solution 5 is better than solution 4 with respect to OB_2. Same thing happens for solutions 3 and 5. Therefore, the Pareto front is made of solutions 3, 4 and 5.

Ensemble learning can be seen as a vote based problem. Suppose that one has a total number of N classifiers $\{C_1, C_2, ..., C_N\}$ trained for a M class problem. Then, the vote based classifier ensemble problem can be defined as finding the combination of votes V per classifier C_i, which will optimize a quality function $F(V)$. V can either represent a binary matrix (binary vote based ensemble) or a matrix containing real values (real/weighted vote based ensemble) of size $N \times M$. In case of binary voting, $V(i,j)$ represents whether C_i is permitted to vote for class M_j. $V(i,j) = 1$ is interpreted as the i^{th} classifier is permitted to vote for the j^{th} class else $V(i,j) = 0$ is interpreted as the i^{th} classifier is not permitted to vote for the j^{th} class. In case of real voting, $V(i,j) \in [0,1]$ quantifies the weight of vote of C_i for the class M_j. If a particular classifier is confident in determining a particular class, then more

weight should be assigned for that particular pair, otherwise less weight should be attributed.

In terms of MOO formulation, the classifier ensemble problem at hand is defined as determining the appropriate combination of votes V per classifier such that objectives $O_1(V)$ and $O_2(V)$ are simultaneously optimized and O_1 = recall and O_2 = precision.

4.4.2 Evolutionary Procedure

The single and multiobjective based methods to solve both the classifier ensemble problems mentioned above use Genetic Algorithms (GAs) [Goldberg, 2006] as the search technique. The single objective formulations of both the combination techniques are solved using some methods based on the search capabilities of genetic algorithm. Similarly our MOO based solutions to solve the above mentioned classifier ensemble problems are based on the search capabilities of nondominated sorting GA-II [Deb et al., 2002].

String Representation: In order to encode the classifier ensemble selection problem in terms of genetic algorithms, we propose to study three different representations.

(1) Simple Classifier Ensemble (SCE): In order to solve the simple classifier ensemble problem, we have used binary encoding. If the total number of available classifiers is N, then the length of the chromosome is N. As an example, the encoding of a particular chromosome is represented in Figure 4.2. Here, $N = 19$, i.e., total 19 different classifiers are built. The chromosome represents an ensemble of 7 classifiers (first, third, fourth, seventh, tenth, eleventh and twelfth classifiers).

The entries of each chromosome are randomly initialized to either 0 or 1. Here, if the i^{th} position of a chromosome is 0 then it represents that i^{th}classifier does not participate in the classifier ensemble. Else, the value 1 designates that the i^{th}

classifier participates in the classifier ensemble. If the population size is P then all the P number of chromosomes of this population are initialized in the above way.

(2) Binary Vote based Classifier Ensemble (BVCE): Each individual classifier is allowed to vote or not for a specific class M_j. The chromosome is of length $N \times M$ and each position takes either 1 or 0 as value. As an example, the encoding of a particular chromosome is represented in Figure 4.3. Here, $M = 3$ and $O = 4$ (i.e., total 12 votes can be possible). The chromosome represents the following voting combination.

Classifier 1 is allowed to vote for classes 1 and 4;
Classifier 2 is allowed to vote for classes 1 and 2;
Classifier 3 is allowed to vote for classes 2, 3 and 4.

The entries of each chromosome are randomly initialized to either 0 or 1. Here, if the i^{th} position of a chromosome is 0 then it represents that $(i/4 + 1)^{th}$ classifier is not allowed to vote for the $(i \bmod 4)^{th}$ class. Else, if it is 1 then it means that $(i/4 + 1)^{th}$ classifier is allowed to vote for the $(i \bmod 4)^{th}$ class. If the population size is P then all the P number of chromosomes of this population are initialized in the above way.

(3) Real/weighted Vote based Classifier Ensemble (RVCE): all classifiers are allowed to vote for a specific class M_j with a different weight for each class. The chromosome is of length $N \times M$ and each position takes a real value. As an example, the encoding of a particular chromosome is represented in Figure 4.4. Here, $M = 3$ and $O = 3$ (i.e., total 9 votes can be possible). The chromosome represents the following voting combination:

The weights of votes for 3 different output classes for classifier 1 are 0.59, 0.12 and 0.56, respectively. Similarly, weights of votes for 3 different output classes are 0.09, 0.91 and 0.02, respectively for classifier 2 and 0.76, 0.5 and 0.21, respectively for classifier 3.

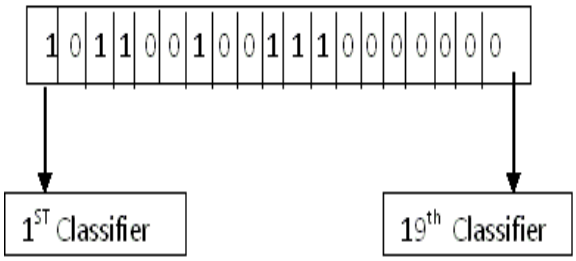

Figure 4.2: Chromosome Representation for Solving the Simple Classifier Ensemble Selection Problem

Here, we use real encoding, i.e. the entries of each chromosome are randomly initialized to a real value (r) between 0 and 1. Here, $r = \frac{rand()}{RAND\_MAX+1}$. If the population size is P then all the P number of chromosomes of this population are initialized in the above way.

Fitness: Each individual chromosome corresponds to a possible ensemble solution V, which must be evaluated in terms of fitness. Let the number of available classifiers be N and their respective individual F-measure values by class F_{ij}, $i = 1 \ldots N, j = 1 \ldots M$ (i.e. F_{ij} is the F-measure of C_i for class M_j). For a given query q, receiving class M_j is weighted as in Equation 4.1 where the output class assigned by C_i to q is given by $op(q, C_i)$. Note that in the case of SCE, $V(i, j)$ is redefined as $V(i, .)$ and F_{ij} as $F_{i.}$.

$$f(q, M_j) = \sum_{i=1:N \& op(q,C_i)=M_j} V(i, j) \times F_{ij}. \tag{4.1}$$

Finally, the class of the query q is given by $argmax_{M_j} f(q, M_j)$. As such, classifying all queries from a development set gives rise to two fitness (or objective)

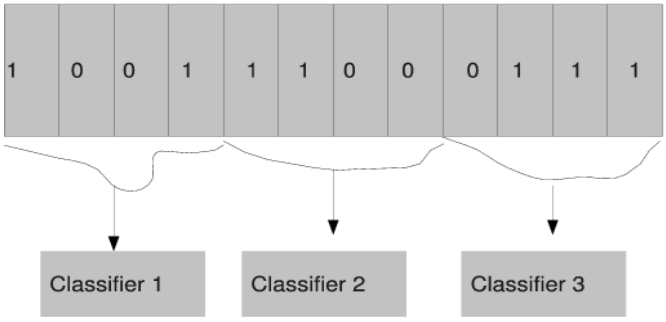

Figure 4.3: Chromosome Representation for Binary Vote Based Classifier Ensemble Selection Problem

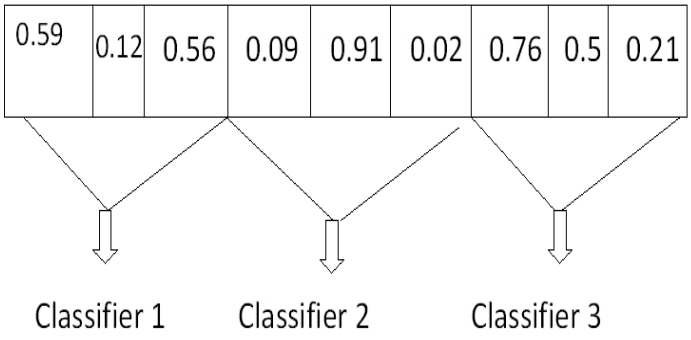

Figure 4.4: Chromosome Representation for Solving the Weighted Vote Based Classifier Ensemble Selection Problem

values, which are respectively recall (O_1) and precision (O_2) and must be optimized simultaneously.

Optimization and Selection: The multi-objective optimization problem is solved by using the Non-dominated Sorting Genetic Algorithm (NSGA-II) [Deb et al., 2002]. The most important component of NSGA-II is its elitism operation, where the non-dominated solutions present in the parent and child populations are moved to the next generation. The chromosomes present in the final population provide the set of different solutions to the ensemble problem and represent the Pareto optimal front.

It is important to note that all the solutions are important, representing a different way of ensembling the set of classifiers. But for the purpose of comparison with other methods, a single solution is required to be selected. For that purpose, we choose the solution that maximizes the harmonic mean of precision and recall i.e. F-measure based on its optimized sub-parts recall and precision as shown in equation 4.2.

$$F\text{-measure} = \frac{2 \times recall \times precision}{recall + precision}. \tag{4.2}$$

4.5 Experiments

Experiments are run over a two-step process. First, $N = 28$ individual classifiers are learned using 10-fold cross validation[6] over a subset of 80 training instances (20 examples for each of the $M = 4$ classes) randomly selected from the initial training set of NTCIR-TQIC containing 100 queries. For each classifier C_i, $F_{i.}$ (global F-measure) and F_{ij} (F-measure for class M_j) values are stored. All experiments were run over the Weka platform[7] with default parameters. Following Weka's denomination, the list of the 28 classifiers is as follows: NaiveBayes, NBTree, NNge, AdaBoostM1, Bagging, BayesNet, BFTree, ClassificationViaRegression, DecisionTable, FT, J48, JRip, IB1,

[6]Note that cross-validation is already an ensemble technique.

[7]`http://www.cs.waikato.ac.nz/ml/weka/`

IBk, Kstar, LWL, LMT, Logistic, LogitBoost, MultiBoostAB, MultilayerPerceptron, RandomCommittee, RandomForest, RBFNetwork, REPTree, RotationForest, SimpleLogistics and SMO. In order to assess the quality of each individual classifier, each one was tested on the NTCIR-TQIC test data set containing 300 unseen queries (75 for each class). Results of the top 5 classifiers are given in Table 4.3.

Table 4.3 Results of single learning strategies.

Classifiers	Precision	Recall	F-measure
Logistic	**82.9**	**75.0**	**78.8**
RandomForest	82.6	70.0	75.8
RotationForest	77.5	70.0	73.6
LMT	69.2	65.0	67.0
SimpleLogistics	69.2	65.0	67.0

The second step of the experiment is the optimization procedure. For that purpose, the remaining 20 query examples (5 for each class) from the NTCIR-TQIC training data set are used. We call it the development set. Based on the development set, the evolutionary optimization using NGSA-II is run for three representations (SCE, BVCE, RVCE) and the best solution is selected based on maximum F-measure as defined in equation 4.2. Performance results are presented in Table 4.4 and compared to two baselines ensemble techniques (BSL1, BSL2). BSL1 corresponds to Boosting with the single Logistic classifier and BSL2 is a SVM solution with 28 features each one corresponding to the output class (i.e. past, recency, future, atemporal) of each of the 28 classifiers.

Table 4.4 Results of ensemble learning strategies.

Measures	RVCE	BVCE	SCE	BSL1	BSL2
Precision	**92.2**	85.1	86.0	82.9	77.5
Recall	**90.0**	85.0	83.7	75.0	75.0
F-measure	**91.1**	85.0	84.8	78.7	76.2

As expected, our methodology outperforms BSL1 by 12.4% and BSL2 by 14.9% in terms of F-measure for the RVCE representation. In particular, BSL1 suffers from the use of a single classifier family while BSL2 can not generalize over the small amount of training data (only 20 examples). Moreover, the most fine tuned strategy in terms of ensemble learning evidences improved results when compared to coarse-grain solutions. Improvements of 6.1% and 6.3% are respectively shown against BVCE and SCE.

In order to understand the spectrum of the different solutions on the Pareto front, we present in Table 4.5 three different situations: the solution that maximizes precision (line 1), the solution that maximizes recall (line 2) and the solution that maximizes F-measure (line 3). Results show that high overall performances are provided by every solution. But, depending on the application at hand, one may expect to find a better tuned configuration.

Table 4.5 Precision and recall spectrum.

Measures	Recall	Precision	F-measure
Max precision	88.7	**93.1**	90.9
Max recall	**90.8**	90.8	90.8
Max F-measure	90.0	92.2	**91.1**

Finally, comparative accuracy results are given against state-of-the-art solutions from NTCIR-11 Temporalia TQIC in Table 4.6[8]. Our solution evidences highly improved results overall as well as for each individual class. In particular, accuracy improvements of 10% for past, 19% for recency, 9% for future, 10% for atemporal and 16% overall are achieved against best existing studies.

[8]Results are taken from [Joho et al., 2014b].

Table 4.6 Comparative accuracy results to state-of-the-art techniques presented in NTCIR-11 Temporalia task.

Classes	RVCE	#1 [Yu et al., 2014]	#2 [Shah et al., 2014]	#3 [Hou et al., 2014]
Past	**0.95**	0.85	0.75	0.79
Recency	**0.82**	0.48	0.56	0.63
Future	**0.94**	0.85	0.81	0.64
Atemporal	**0.89**	0.77	0.79	0.71
All	**0.90**	0.74	0.73	0.69

4.6 Feature Importance Evaluation

In order to better assess the importance of each individual feature, we used Information Gain (IG) attribute evaluation method. IG measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution.

$$\text{InfoGain(Class,Attribute)} = \text{H(Class)} - \text{H(Class |Attribute)}$$

where H is the information entropy.

Top five (5) features with their merits are presented in Table 4.7.

Table 4.7 Top five (5) informative features

Feature	IG	Main resource/tool
D_b_Dates	0.245	GTE [Campos et al., 2012a]
N_o_FW	0.219	TempoWordNet
N_o_FS	0.135	STC
C_o_Q	0.130	STC
C_o_Date	0.114	GTE [Campos et al., 2012a]

The result clearly evidences that TempoWordNet embodies enough time-related information that is beneficial for TQIC task. Results also show that both the extra collected data i.e. Bing web snippets and GTE year dates are compensating the data scarcity problem here.

4.7 Summary

Web search is strongly influenced by time. Understanding search intent behind a user's query is a crucial part towards improving the performance of information access system. For instance it can be useful before applying a suitable retrieval and ranking model. In the temporal search domain, a system should be able to detect a query that contains an underlying temporal information need. TQIC task aims at determining the temporal orientation of the user's information need, i.e., whether she is interested in information from/about the past, present, future or atemporal.

The TQIC task is challenging mainly because there is little information associated with a web search query. User's usually summarize their information needs in the form a query which composed of only 3-5 keywords and the time when the query was issued. Apart from that ambiguity, some queries may lead to different interpretations. Also, it limits the applicability of certain methods. For example, exiting Word Sense Disambiguation (WSD) methods usually do not perform well on short strings.

We examine how the temporal knowledge embedded in TempoWordNet can be useful in temporal query intent classification scenario. We consider the task as a machine learning classification problem. Due to the small amount of gold training data, multi-objective based ensemble learning is applied, whose underlying idea is to reduce bias by combining multiple classifiers instead of relying on a single one. For our purpose, we propose a set of features which can easily be extracted from different freely available resources. In particular, we use TempoWordNet and STC obtained at the time of building TempoWordNet to design several features. Analysis of the proposed features shows that features designed from TempoWordNet and STC play crucial role in the classification process.

Chapter 5

Improvement on TempoWordNet

Previously in Chapter 3, different methods are used to strategically expand temporality inside WordNet that serve as central point in the building process of TWnL, TWnP, TWnH respectively. The main motivation to experiment several methodologies is to propose more reliable TWn(s). However, intrinsic and extrinsic evaluation results in the form of human-annotator agreement and sentence temporal classifications witness to be on the fence. In this chapter, we introduce an iterative strategy that temporally extends glosses based on some version of TempoWordNet to obtain a potentially more reliable TempoWordNet. The process is iterated until some toping criteria is reached.

This chapter is organized as follows. The next section introduces the motivation and objectives. Section 5.2 presents our overall methodology. Comparative features of different TempoWordNet versions are also covered in the same section. Intrinsic and extrinsic evaluation of the propose iterative process is covered in Section 5.3.

5.1 Motivation and Objective

In order to be of service to the temporal research community, TempoWordNet is hosted in `https://tempowordnet.greyc.fr/`, that gives free access of TWnL, TWnP, and TWnH to the community. Since its public release in March 2014, we

received mixed feedback from the community and encouraging traffic to the website. As of this writing (30.07.2015), the resource is downloaded for more than 5800 times. A snapshot of the traffic details in terms of countries is given below:

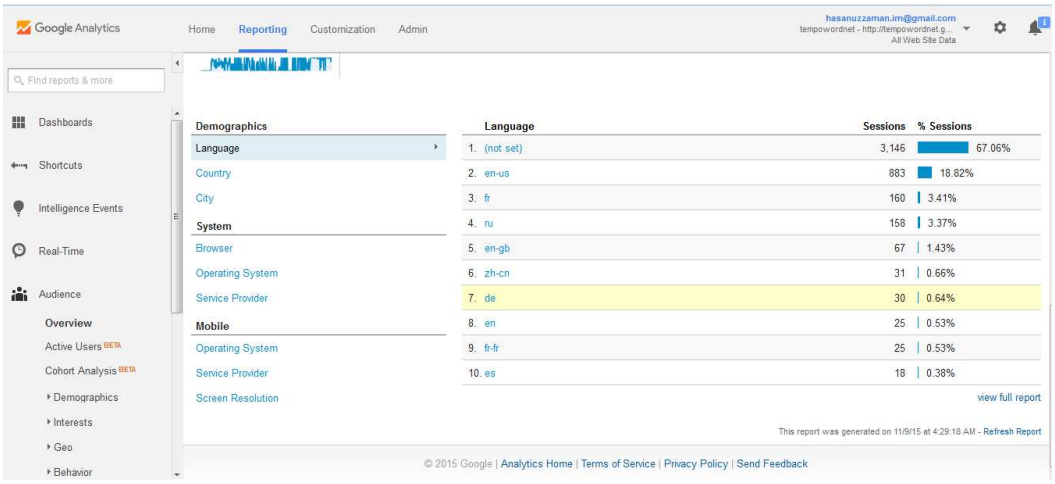

Figure 5.1: Traffic details of TWn hosting website

In the previous chapters, we showed that notable improvement can be achieved for sentence temporal classification task by augmenting the sentence with temporal knowledge from TWn. We have also examined how a search application can benefit from the temporal information embodied in TWn. Precisely, TWn is tested within the context of the TQIC task of NTCIR-11 Temporalia, where the main goal is to predict the underlying temporal intent (past, recency, future or atemporal) of a given search engine query. In addition, we showed that TWn can be useful to time-tag web snippets.

However, use of TWn shows mixed results which motivate the construction of a more reliable resource. For example, classification results for the annotated sentences that come with the Negex algorithm[1] is shown promising in Salmon Run blog[2], while

[1]https://code.google.com/p/negex/
[2]http://sujitpal.blogspot.fr/2014/04/find-temporalaty-of-sentences-with.html

less comprehensive results are shown in [Filannino and Nenadic, 2014], where TWn learning features did not lead to any classification improvements.

Therefore, we experiment an iterative strategy to build an accurate TWn both in terms of human judgment and classification results. Our underlying idea is simple. It is based on our previous findings at the time of building TWn that (1) synsets are temporally classified based on their gloss and (2) temporal sentence classification is boosted by TWn, temporally expanding glosses based on a given TempoWordNet at step t (TWn^t) may allow to obtain a more accurate TempoWordNet at step $t+1$ (TWn^{t+1}) when propagating temporal connotations.

The overall strategy is to start from previously selected lists of handcrafted temporal seed synsets and a given propagation strategy, TWn^0 is constructed based on Wn^0 (WordNet at step 0). TWn^0 is then used to temporally expand Wn^0 glosses giving rise to Wn^1. The propagation strategy is then re-run over Wn^1 to give rise to TWn^1 which in turn is used to temporally expand Wn^0 glosses giving rise to Wn^2 and so on and so forth until some stopping criterion.

5.2 Methodology

In order to propose a more reliable temporal lexical resource, we propose to rely on two previous findings. First, automatic synset time-tagging can be achieved by gloss classification with some success. The underlying idea is that the definition of a given concept embodies its potential temporal dimension. Second, temporal sentence classification can be improved when temporal unigrams are augmented with their synonyms stored in TWn version.

From these two assumptions, a straightforward conclusion may be drawn to improve the reliability of the temporal expansion process within WordNet. Indeed, glosses are sentences defining the concept at hand. As a consequence, temporally expanding glosses based on some TWn version may improve the performance of the

classifier used to propagate the temporal connotations and as a consequence meliorate the intrinsic quality of the obtained temporal resource.

Note that this process can be iterated. If a better TWn can be obtained at some step, better gloss expansion may be expected and as a consequence a more accurate temporal resource may be obtained, which in turn may be reused for gloss expansion, and so on and so forth. This iterative strategy is defined in algorithm 2, where the process stops when a stopping criterion is satisfied.

Algorithm 2 Iterative TempoWordNet algorithm

1: $sl \leftarrow$ list of temporal seed synsets
2: $ps \leftarrow$ propagation strategy
3: $i \leftarrow 0$
4: $\text{Wn}^0 \leftarrow \text{Wn}$
5: **repeat**
6: $\quad \text{TWn}^i \leftarrow \text{PropagateTime}(\text{Wn}^i, sl, ps)$
7: $\quad \text{Wn}^{i+1} \leftarrow \text{ExpandGloss}(\text{Wn}, \text{TWn}^i)$
8: $\quad i \leftarrow i + 1$
9: **until** stopping criterion
10: return TWn^i

Let sl be the set of temporal seed synsets, ps a propagation strategy (lexical, probabilistic or hybrid) and a working version of WordNet Wn^0. The iterative construction of TWn is as follows:

Based on sl, an initial two-class (temporal vs atemporal) classifier is learned over a 10-fold cross validation process, where synsets are represented by their gloss constituents (i.e. unigrams) weighted by gloss frequency. The propagation of the temporal connotations is then run based the learned classifier and the given propagation strategy ps. New expanding synsets are included in the initial seeds list and the exact same learning process iterates N times so to ensure convergence in terms of classification accuracy. At the end of this process, the synsets of the working version of WordNet Wn^0 are either tagged as temporal or atemporal depending on

classification probability, thus giving rise to two distinct partitions Wn_t^0 (temporal Wn^0) and $\text{Wn}_{\bar{t}}^0$ (atemporal Wn^0). In order to fine tune Wn_t^0 in past, present and future synsets, the exact same procedure is run exclusively based on the past, present and future synsets from *sl* over Wn_t^0. When classification convergence is reached[3], all synsets from Wn_t^0 are time-tagged as past, present or future giving rise to Wn_{tt}^0. Note that a synset can neither be past, present nor future (e.g. "sunday") although it is temporal. In this case, near equal class probabilities are evidenced. At the end of the construction process, $\text{TWn}^0 = \text{Wn}_{\bar{t}}^0 \cup \text{Wn}_{tt}^0$.

Once TWn^0 has been constructed, it can be used to temporally expand WordNet (Wn) glosses giving rise to the second working version of WordNet Wn^1. In particular, within each gloss of Wn, all temporal words from TWn^0 are searched for. If one temporal concept is found in the gloss, then its synonyms are added to the gloss thus enlarging the possible lexical overlap between temporal glosses. So, each gloss is now represented by its unigrams plus the synonyms of its temporal constituents as a bag of words and is noted Wn^1. Note that word sense disambiguation is performed based on the implementation of the Lesk algorithm [Lesk, 1986] of the NLTK toolkit [Loper and Bird, 2002]. This process refers to line 7 of algorithm 2.

So, from the new Wn^1, the next TempoWordNet version TWn^1 can be processed, which in turn can give rise to a new WordNet working version Wn^2, which will lead to TWn^2 and so on and so forth. This iterative process stops when some stopping criterion is reached. Many ideas can be experimented here. However, we propose that the final TempoWordNet version is obtained when the difference between TWn^i (TempoWordNet at step i) and TWn^{i-1} (TempoWordNet at step $i-1$) is marginal in terms of temporal sets, i.e. $\text{TWn}^i \setminus \text{TWn}^{i-1} \leq \epsilon$, which means that the set of distinct temporal synsets converges.

[3] N iterations are performed.

Table 5.1 Comparative features of different TempoWordNet versions.

TempoWordNet version	TWnL	TWnP	$TWnH^0$	$TWnH^1$	$TWnH^2$	$TWnH^3$
# temporal synsets	21213	53001	17174	2020	2804	2832
# past synsets	1734	2851	2547	305	120	1308
# present synsets	16144	19762	842	1247	2181	765
# future synsets	3335	30388	13785	468	503	759
# atemporal synsets	96402	64614	100441	115639	114855	114827
$TWn^i \setminus TWn^{i-1}$	-	-	-	15154	784	28
Fixed-marginal κ	0.507	0.520	0.420	0.625	0.616	0.599
Free-marginal κ	0.520	0.520	0.440	0.850	0.700	0.774

Table 5.2 F_1-measure results for temporal sentence, tweet and query intent classification with different TempoWordNet versions performed on 10-fold cross validation with SVM with Weka default parameters.

TempoWordNet version	without TWn	TWnL	TWnP	$TWnH^0$	$TWnH^1$	$TWnH^2$	$TWnH^3$
Sentence classification	64.8	66.7	69.3	68.6	68.4	69.7	71.4
Tweet classification	39.7	49.1	51.5	49.8	51.9	52.5	53.1
Query intent classification	75.3	78.0	78.8	75.9	78.3	79.0	80.1

5.2.1 Experimental Setup

As for experimental setups, we used (1) the *sl* seeds list used in our earlier experiments (2) the *ps* hybrid propagation strategy. Note that lexical propagation does not spread over a wide range of concepts inside WordNet semantic network and probabilistic propagation shows semantic shift problems. (3) version 3.0 of WordNet[4] for Wn and (4) $\epsilon = 100$.

[4] `https://wordnet.princeton.edu/`

With respect to the learning procedures, the SVM implementation of Weka[5] was used with default parameters and convergence was ensured by iterating the temporal propagation $N = 50$ times.

5.3 Evaluation

Our methodology is evaluated both intrinsically and extrinsically, the underlying idea being that a reliable resource must evidence high quality time-tagging as well as improved performance for some application.

5.3.1 Intrinsic Evaluation

In order to assess human judgment about the temporal parts of TempoWordNet, inter-rater agreement with multiple raters is performed. Three annotators are presented with 50 temporal synsets and respective glosses, and must decide upon their correct classification i.e. temporal or atemporal. Note that past, present and future connotations are only indicative of the temporal orientation of the synset but cannot be taken as a strict class. Indeed, there are many temporal synsets, which are neither past, present nor future (e.g. "monthly"). The free-marginal multirater kappa [Randolph, 2005] and the fixed-marginal multirater kappa [Siegel and Castellan, 1988] values are reported in Table 5.1.

According to the interpretation of Kappa (κ) values by Landis and Koch [Landis and Koch, 1977] presented in Table5.3, our results show moderate agreement for previous versions of TempoWordNet i.e TWnL obtained by lexico-semantic expansion, TWnP constructed using probabilistic propagation, and TWnH^0 built on following hybrid propagation process. However, substantial agreement is obtained for the successive iterative TWn versions. Table 5.1 also presents figures about the distribution of temporal synsets of each TempoWordNet version. Interestingly, the iterative versions tend to time-tag a much smaller proportion of synsets when com-

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

pared to previous ones. Note that the number of past, present and future synsets is based on the highest probability given by the temporal classifier, which does not necessarily imply that the synset belongs to the given class (e.g. almost equal probabilities can be evidenced).

Table 5.3 Kappa values interpretation.

κ	**Interpretation**
<0	Poor agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

5.3.2 Extrinsic Evaluation

Temporal sentence classification has traditionally been used as the baseline extrinsic evaluation and consists in labeling a given sentence as past, present or future. In order to produce comparative results with prior versions of TWn, we test our methodology in a similar way as earlier on the balanced data set created in Section 3.4.2 of Chapter 3 from the SemEval-2007 corpus.

Moreover, we propose to extend experiments on a corpus of 300 temporal tweets. This corpus contains 100 past, 100 present and 100 future tweets, which have been time-tagged by annotators of the CrowdFlower[6] platform. Recent work in natural language processing suggests that crowdsourcing annotations from the untrained public can provide annotated data at similar annotation quality as expert annotators, but for a fraction of the cost [Sheng et al., 2008]. We chose CrowdFlower as our crowdsourcing platform, as it has access to a large user base (it uses Amazon Mechan-

[6] `http://www.crowdflower.com/`

ical Turk to find workers), but adds an extra validation layer to attempt to address quality concerns, as this has been an issue in many applications of crowdsourcing in natural language annotation tasks [Callison-Burch and Dredze, 2010].

Each annotation instance consists of a single tweet. To prepare a set of unlabeled instances for annotation, tweets are collected using the Twitter streaming API[7], which offers a live stream of messages posted on Twitter. Unwanted symbols and characters are removed from the tweets. Annotators were asked to choose from 4 (four) possible categories such as *Past, Present, Future,* and *None of these* based on the underlying temporal orientation of tweets. An additional choice *None of these* is included to allow annotators to indicate ambiguous (tweets with unclear or multiple temporal orientations) and/or atemporal tweets.

We required that at least 3 judgments be made for each annotation, and majority voting is then used to decide on a final label for each annotation. To ensure the quality of the judgments we have obtained, we make use of a validation facility provided by CrowdFlower. Pre-annotated gold instances are mixed together with unlabeled instances. These gold instances are used to validate the annotations made by each annotator. Annotators were not informed which were the gold instances. During the annotation process, annotators who failed to label these gold instances correctly were stopped from proceeding with the task, and the annotations they made are discarded. Some examples are given as follows:

- So windy last night the tiles blew off the roof llf. (**Past**),
- Today we commence part 43384916 on the master plan. (**Present**),
- What's the homework for tomorrow, tomorrow is saturday oh yeah. (**Future**)
- Why i love new year's eve. (**None of these**).

With regard to the representation, each tweet is represented by a feature vector where each attribute is either a unigram or a synonym of any temporal word con-

[7] `https://dev.twitter.com/streaming/overview`. Last accessed on 02-05-2015

tained in the tweet and its value is tf.idf. Stopwords removal is performed using Weka so to better access the benefits of TempoWordNet. Word sense disambiguation is performed in order to choose exact concept of a word. The results of our experiments are presented in Table 5.2.

In order to strengthen comparative evaluation, we also propose to tackle the TQIC task of NTCIR-11 Temporalia [Joho et al., 2014b], where a given search engine query must be tagged as past, recency, future or atemporal. For that purpose, we use the 400 queries provided by the organizers, which are equally distributed by temporal class. In particular, a query is represented in the same way as sentences and tweets plus the additional time-gapped feature, which compares the issue date of the query and the most confident year retrieved either from the query itself or from its web snippets results. It is important to mention that stopwords are not removed as queries are small. Comparative classification results are reported in Table 5.2.

For all experiments, TWnH^3 produces highest classification results with respectively 2.1%, 1.6% and 1.3% improvements for sentence, tweet and query temporal classification over the second best (non-iterative) TempoWordNet version. Note that all improvements are statistically relevant and steadily occur between TWnH^0, TWnH^1, TWnH^2 and TWnH^3.

5.4 Summary

We proposed an iterative strategy to produce more reliable TempoWordNet. The underlying idea is that based on the temporal expansion of glosses with some version of TempoWordNet at step t, a more accurate resource can be obtained at step $t + 1$. Manual evaluation result showed substantial agreement for the successive iterative TWn versions compared to moderate agreement for previous versions of TempoWordNet (TWnL, TWnP, TWnH). Extrinsic evaluation evidences improved sentence classification results when compared to previous versions of TempoWordNet.

Moreover, to strengthen comparative evaluation, we proposed to test the iterative strategy in the context of tweet and query temporal classification. For the first task, we created a gold standard corpus consists of 300 tweets with the help of a crowdsourcing task. For the second one, we used gold standard NTCIR TQIC dataset. Iterative TempoWordNet (TWnH^3) produced highest classification results for the both tasks.

Chapter 6

Conclusions and Future Directions

This dissertation focused on the investigation and understanding of the different ways time-sensitive concepts/expressions are conveyed in natural language, on the implementation of different approaches to identify these concepts/expressions and arrange them in a large semantic network, on the analysis of errors and challenges faced during implementation process, on the extensive evaluations of each approach, on the implementation of a framework to understand temporal intent behind users query inspired from the results of this investigation, and on the qualitative improvement of the resources.

The work presented in this thesis is a piece of research in language engineering. Therefore, the main stress was to build a temporal resource to understand the language of time. The main requirements of the resource were to be reliable and beneficial for larger NLP and IR application.

In designing the methodologies, the main aim was to ensure their reproducibility and wide applicability.

The ability to capture the time information conveyed in natural language is essential to many natural language processing applications such as information retrieval, question answering, and automatic summarization. Associating word senses with

temporal dimensions, namely, *past, present, future,* and *atemporal*, to grasp the temporal information in language is relatively straightforward task for humans by using world knowledge. A lexical temporal resource associating word senses with temporality is crucial for the computational tasks aiming at interpretation of language of time in text. This thesis first addressed the building of a lexical temporal resource by automatically time-tagging each synset of WordNet by iteratively learning temporal classifiers from an initial set of time-sensitive synsets and a given propagation strategy. As such, each synset was automatically time-tagged with four dimensions i.e. *atemporal*, *past*, *present* and *future*, led to the different TempoWordNets depending on the propagation strategy.

Although the process of time tagging WordNet has been inspired by the initial idea of [Esuli and Sebastiani, 2005], it can not be compared to the one of opinion tagging. The subjective information linked to a word is both about connotation and denotation. Thus, hyponymy is a privileged relation for propagating opinions. Indeed, both the main sense and the same semantic orientation are kept along the hyponymy relation. On the contrary, we observed that the temporal information is more associated to denotation than connotation. As such, although hyponyms of a given temporal synset still have the same denotation, the temporal connotation may be lost. As a consequence, enhanced propagation strategies (probabilistic and hybrid), instead of just lexico-semantic relations were proposed in order to improve the intrinsic quality of TempoWordNet.

TempoWordNets are evaluated both manually and automatically. In order to intrinsically evaluate the time-tagged WordNets (TempoWordNets), statistically significant samples of automatically time-tagged WordNet synsets along with their glosses were presented to 3 annotators to measure the multirater agreement values namely free-marginal multirater kappa and the fixed-marginal multirater kappa values. Overall results showed moderate agreement for the three TempoWordNets: TWnL (lexico-semantic expansion), TWnP (probabilistic expansion) and TWnH (hybrid expansion). Inter-annotator agreement values revealed the difficulty of the task

for humans as they do not agree on a great deal of decisions. This is particularly due to the fact that the temporal dimension of synsets is judged upon their glosses and not directly on their inherent concept. For example, "dinosaur" can be classified as *temporal* or *atemporal* as its gloss *any of numerous extinct terrestrial reptiles of the Mesozoic era* allows both interpretations.

In order to evaluate TempoWordNet automatically, we proposed to evidence its usefulness based on an external task: sentence temporal classification. The underlying objective was that a temporal knowledge embedded in TempoWordNet(s) can help to classify sentences into three different categories: *past*, *present* and *future*. The experiments made for sentence temporal classification showed that TempoWordNet allows 13.9% improvements of F_1-measure against the vector space model representation and 14.7% against the semantic vector space model obtained with the existing WordNet time subtree. We also evidenced the importance played by stop words in sentence temporal classification where improvements with TempoWordNet are less expressive (2.2%).

Temporal information embedded in web search queries offer an interesting means to further enhance the functionality of current information access systems. Understanding such temporal orientation and utilizing it in web search scenarios can significantly improve the current functionality of information access systems.

To substantiate our claim, we examined the usefulness of TempoWordNet for this task and proposed a framework to determine the temporal intent behind a user's search query. We tackled the problem of identifying temporal intent of queries from a machine learning point of view. Due to the small amount of gold training data, we proposed an ensemble learning solution, whose underlying idea is to reduce bias by combining multiple classifiers instead of relying on a single one. In particular, recently developed multi-objective based ensemble techniques have been applied that (1) allows to accurately classify queries along their temporal intent and (2) identifies a set of performing solutions thus offering a wide range of possible applications. For

our purpose, we made use of a set of features which can easily be extracted from different freely available resources. Experiments showed that correct representation of the problem can lead to great classification improvements when compared to recent state-of-the-art solutions and baseline ensemble techniques.

Improvements have been experienced for sentence temporal classification task by augmenting the sentence with temporal knowledge from TempoWordNet(TWnL, TWnP, TWnH). A search application also got benefited from the temporal information embodied in TWn. Precisely, TWn is tested within the context of the TQIC task of NTCIR-11 Temporalia, where the main goal is to predict the underlying temporal intent (past, recency, future or atemporal) of a given search engine query. In addition, we showed that TWn can be useful to time-tag web snippets. However, it was a hard task to choose between the different TempoWordNet versions. For example, the intrinsic evaluation tends to mention that TWnH is preferred by human annotators while TWnP evidences best results for temporal sentence classification. This motivated us for the construction of a more reliable resource.

Based on our previous findings at the time of building TWn that (1) synsets are temporally classified based on their gloss and (2) temporal sentence classification is boosted by TWn, an easy conclusion was made that temporally expanding glosses based on a given TempoWordNet at step t (TWn^t) should allow to obtain a more accurate TempoWordNet at step $t+1$ (TWn^{t+1}) when propagating temporal connotations. So, we proposed an iterative strategy that temporally extends glosses based on TWn^t to obtain a potentially more reliable TWn^{t+1}. Intrinsic and extrinsic evaluations evidenced improved results when compared to previous versions of TempoWordNet.

In light of the above, we deeply believe that our contribution towards building a reliable temporal resource will attract the community to consider the temporal knowledge conveyed by the resource as a feature for various time-related NLP and IR applications. As a consequence, we provide free access to this resource as well

as all developing materials at http://tempowordnet.greyc.fr. The resource can also be found on the page provides access to wordnets in a variety of languages at http://compling.hss.ntu.edu.sg/omw/.

6.1 Future Work

In this subsection we outline a number of extensions to our research involve specific approaches that would improve the quality of the temporal lexical resource described in this thesis, evaluation technique, and applications that would use this resource to address NLP and IR problems.

For the first category, one possible line of research would be to experiment semi-supervised graph classification algorithm build on an optimization theory namely the max-flow min-cut theorem for associating temporality to word senses. Underlying idea behind classification with minimum cuts (Mincuts) in graph is that similar items should be grouped together in the same cut. All items in the training/test data can be seen as vertices in a graph with undirected weighted edges between them. Each edge weight corresponds to either one of the two information. The first one, *individual score* is the non-negative estimates of each vertex's preference for being in a particular class based on the features of that vertex alone. While the later one, *association scores* represents a non-negative estimates of how important it is that two different vertices be in the same class.

Formulating the task of temporality detection problem on word senses in terms of graphs would allows us to model item-specific and pair-wise information independently. Therefore, it is a very flexible paradigm where we could have different views on the data. For example, rule based approach or machine learning algorithms employing linguistic and other features representing temporal indicators can be used to derive *individual scores* for a particular item in isolation. The edges weighted by the *individual scores* of a vertex (=word sense) to the temporal/atemporal can be interpretative as the probability of a word sense being temporal or atemporal with-

out taking similarity to other senses into account. And we could also simultaneously use conceptual-semantic and lexical relations from WordNet to derive the *association scores*. The edges between two items weighted by the *association scores* can indicate how similar/different two senses are.

One specific task that could be investigated further is to explore the effect of other graph construction methods using freely available online dictionaries including thesaurus and distributional similarity measures.

Another aspect that needs further attention is to propose an evaluation benchmark to evaluate the performance of the temporal classification and the quality of TempoWordNet. The evaluation phase requires a gold standard data to be able to conduct the assessment. Since to our best knowledge there is no resource with temporal associations of word senses (synsets). Therefore, we would like to first design our own annotation task to create a gold standard with the help of a crowdsourcing task. Based on the annotation results of our crowdsourcing task, we are interested to propose an evaluation technique by considering that a synset might be associated with more than one temporal categories. In particular, we will adopt similar evaluation measures of SemEval-2007 English Lexical Substitution Task proposed in [McCarthy and Navigli, 2007], where a system generates one or more possible substitutions for a target word in a sentence preserving its meaning.

Multilingual information access is critical for the acquisition, dissemination, exchange, and understanding of knowledge in the global information society. The accelerated growth in the size, content and reach of Internet, the diversity of user demographics and the skew in the availability of information across languages, all point to the increasingly critical need for multilingual information access. With this in mind, multilingual TempoWordNet is another research area that could be pursued further.

From a resource point of view, when looking at applications that would benefit from TempoWordNet developed in this work, the possibilities are endless.

One research direction that would benefit the research community is temporal re-ranking of web search results. Temporal re-ranking has proved to lead to improve results when queries with an implicit temporal intent are issued [Campos et al., 2014b]. However, all existing models do not take into account the different facets of time and documents are ranked higher if the are temporally relevant. Temporal queries may have different temporal intents such as past, present or future. As such, documents should not be re-ranked only on their general temporal nature but should be treated according to their specific temporal dimensions. Therefore, temporal intent of a given query prior to re-ranking would be incorporated. In particular, the re-ranking function must take into account the discovered intent.

Another application that would benefit from the resource we provided, would be automatic analysis of time-oriented clinical narratives. For example, it should be possible for a system to automatically analyze medical discharge summaries including previous diseases related to the current conditions, treatments, and the family history for medical decision making, data modeling and biomedical research.

Another application of TempoWordNet would be to identify *future focused* popular topics from Twitter. Twitter, one of the fastest growing and most popular micro-blogging services, recently attracts special attention among researchers, since information disseminated through this service is faster than traditional sources. Twitter evidences collective attention across diverse topics from social, economic, health related to the latest local and global news and events. Knowledge of *past* and *present* phenomenon are useful but future-linked popular topics can unprecedentedly support organizations and individuals to detect emerging issues and prepare for otherwise surprising developments. Though, future is uncertain, spotting future oriented popular topic is of utmost interest and importance to business and administrative decision makers as quickly as possible, as it can buy extra precious time for them to make informed decisions. For example, an upcoming product popularity can help to promote business, sales prediction and sensing of future consumer behavior, in turn can

help business decision makers in budget planning, marketing campaigns and resource allocation.

Another line of research that could be pursued to integrate TempoWordNet in the temporal processing phase of a QA system. To this end, temporal knowledge embedded in the resource developed in this work would have to be applied at all the stages involved in the QA process, i.e. Question Processing, Paragraph Retrieval, and Answer Extraction. The methodology that would guide the integration of the resource in a QA system offers a long-term research direction.

In summary, exciting and promising research awaits to unlock the value of temporal lexical resource in many possible applications.

Bibliography

[Aji et al., 2010] Aji, A., Wang, Y., Agichtein, E., and Gabrilovich, E. (2010). Using the past to score the present: Extending term weighting models through revision history analysis. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 629–638, New York, NY, USA. ACM.

[Alonso et al., 2007a] Alonso, O., Baeza-Yates, R., and Gertz, M. (2007a). Exploratory search using timelines. In *SIGCHI 2007 Workshop on Exploratory Search and HCI Workshop*, number 1. Citeseer.

[Alonso et al., 2009a] Alonso, O., Baeza-Yates, R., and Gertz, M. (2009a). Effectiveness of temporal snippets. In *WSSP Workshop at the World Wide Web ConferenceâĂŤWWW*, volume 9.

[Alonso and Gertz, 2006a] Alonso, O. and Gertz, M. (2006a). Clustering of search results using temporal attributes. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 597–598, New York, NY, USA. ACM.

[Alonso and Gertz, 2006b] Alonso, O. and Gertz, M. (2006b). Clustering of search results using temporal attributes. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 597–598. ACM.

[Alonso et al., 2007b] Alonso, O., Gertz, M., and Baeza-Yates, R. (2007b). On the value of temporal information in information retrieval. In *ACM SIGIR Forum*, volume 41, pages 35–41. ACM.

[Alonso et al., 2009b] Alonso, O., Gertz, M., and Baeza-Yates, R. (2009b). Clustering and exploring search results using timeline constructions. In *Proceedings*

of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, pages 97–106, New York, NY, USA. ACM.

[Alonso et al., 2009c] Alonso, O., Gertz, M., and Baeza-Yates, R. (2009c). Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 97–106. ACM.

[Alonso et al., 2011a] Alonso, O., Gertz, M., and Baeza-Yates, R. (2011a). Enhancing document snippets using temporal information. In *String Processing and Information Retrieval*, pages 26–31. Springer.

[Alonso et al., 2011b] Alonso, O., Strötgen, J., Baeza-Yates, R., and Gertz, M. (2011b). Temporal information retrieval: Challenges and opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW)*, pages 1–8.

[BAEZA and Ribeiro-Neto, 2011] BAEZA, Y. and Ribeiro-Neto, B. (2011). Modern information retrieval-the concepts and technology behind search.

[Baeza-Yates, 2005] Baeza-Yates, R. (2005). Searching the future. In *SIGIR Workshop MF/IR*.

[Belkin and Croft, 1992] Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38.

[Biber et al., 1999] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman grammar of spoken and written English*, volume 2. MIT Press.

[Bush and Think, 1945] Bush, V. and Think, A. W. M. (1945). The atlantic monthly. *As we may think*, 176(1):101–108.

[Callison-Burch and Dredze, 2010] Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

[Campos et al., 2011a] Campos, R., Dias, G., and Jorge, A. (2011a). An exploratory study on the impact of temporal features on the classification and clustering of future-related web documents. In *Proceedings of the 15th Portugese Conference on Progress in Artificial Intelligence*, EPIA'11, pages 581–596, Berlin, Heidelberg. Springer-Verlag.

[Campos et al., 2012a] Campos, R., Dias, G., Jorge, A., and Nunes, C. (2012a). Gte: A distributional second-order co-occurrence approach to improve the identification of top relevant dates in web snippets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2035–2039.

[Campos et al., 2009] Campos, R., Dias, G., and Jorge, A. M. (2009). Disambiguating web search results by topic and temporal clustering-a proposal. In *KDIR*, pages 292–296.

[Campos et al., 2014a] Campos, R., Dias, G., Jorge, A. M., and Jatowt, A. (2014a). Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):15.

[Campos et al., 2012b] Campos, R., Dias, G., Jorge, A. M., and Nunes, C. (2012b). Enriching temporal query understanding through date identification: How to tag implicit temporal queries? In *Proceedings of the 2Nd Temporal Web Analytics Workshop*, TempWeb '12, pages 41–48, New York, NY, USA. ACM.

[Campos et al., 2014b] Campos, R., Dias, G., Jorge, A. M., and Nunes, C. (2014b). Gte-rank: Searching for implicit temporal query results. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, pages 2081–2083.

[Campos et al., 2011b] Campos, R., Jorge, A., and Dias, G. (2011b). Using web snippets and query-logs to measure implicit temporal intents in queries. In *SIGIR 2011 Workshop on Query Representation and Understanding*.

[Campos et al., 2012c] Campos, R., Jorge, A. M., Dias, G., and Nunes, C. (2012c). Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 1–8, Washington, DC, USA. IEEE Computer Society.

[Chang and Manning, 2012] Chang, A. X. and Manning, C. D. (2012). Sutime: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.

[Chang et al., 2012] Chang, P.-T., Huang, Y.-C., Yang, C.-L., Lin, S.-D., and Cheng, P.-J. (2012). Learning-based time-sensitive re-ranking for web search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1101–1102, New York, NY, USA. ACM.

[Cleverdon, 1967] Cleverdon, C. (1967). The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd.

[Collins and Quillian, 1969] Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247.

[Costa et al., 2014] Costa, M., Couto, F., and Silva, M. (2014). Learning temporal-dependent ranking models. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 757–766, New York, NY, USA. ACM.

[Cowie and Wilks, 2000] Cowie, J. and Wilks, Y. (2000). Handbook of natural language processing. chapter information extraction.

[Croft and Harper, 1979] Croft, W. B. and Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35(4):285–295.

[Dakka et al., 2008] Dakka, W., Gravano, L., and Ipeirotis, P. G. (2008). Answering general time sensitive queries. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1437–1438, New York, NY, USA. ACM.

[Deb et al., 2002] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):181–197.

[Derczynski et al., 2013] Derczynski, L. R., Yang, B., and Jensen, C. S. (2013). Towards context-aware search and analysis on social media data. In *Proceedings of the 16th international conference on extending database technology*, pages 137–142. ACM.

[Dias et al., 2011] Dias, G., Campos, R., and Jorge, A. M. (2011). Future retrieval: What does the future talk about? In *Workshop on Enriching Information Retrieval of the 34th ACM Annual SIGIR Conference (SIGIR 2011)*, pages 3–pages.

[Dias et al., 2012] Dias, G., Moreno, J. G., Jatowt, A., and Campos, R. (2012). Temporal web image retrieval. In *String Processing and Information Retrieval*, pages 199–204. Springer.

[Diaz, 2009] Diaz, F. (2009). Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 182–191, New York, NY, USA. ACM.

[Dietterich, 2000] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer.

[Dong et al., 2010a] Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., Buchner, K., Liao, C., and Diaz, F. (2010a). Towards recency ranking in web search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 11–20, New York, NY, USA. ACM.

[Dong et al., 2010b] Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. (2010b). Time is of the essence: Improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 331–340, New York, NY, USA. ACM.

[Esuli and Sebastiani, 2005] Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 617–624.

[Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, pages 417–422.

[Fellbaum, 1998a] Fellbaum, C. (1998a). *WordNet*. Wiley Online Library.

[Fellbaum, 1998b] Fellbaum, C. (1998b). *WordNet: An Electronic Lexical Database*. Bradford Books.

[Filannino and Nenadic, 2014] Filannino, M. and Nenadic, G. (2014). Using machine learning to predict temporal orientation of search engines' queries in the temporalia challenge. In *NTCIR-11 Conference (NTCIR)*, pages 438–442.

[Fleiss, 1971] Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[Gangemi et al., 2003] Gangemi, A., Navigli, R., and Velardi, P. (2003). The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE*, pages 820–838. Springer.

[Goldberg, 2006] Goldberg, D. E. (2006). *Genetic algorithms*. Pearson Education India.

[Gómez-Pérez et al., 2003] Gómez-Pérez, A., Manzano-Macho, D., et al. (2003). A survey of ontology learning methods and techniques. *OntoWeb Deliverable D*, 1(5).

[Greengrass, 2000] Greengrass, E. (2000). Information retrieval: A survey.

[Gruber, 1995] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5):907–928.

[Harman, 1993] Harman, D. (1993). Overview of the first text retrieval conference. In *NATIONAL ONLINE MEETING*, volume 14, pages 181–181. LEARNED INFORMATION (EUROPE) LTD.

[Hotho et al., 2005] Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62.

[Hou et al., 2014] Hou, Y., Chen, Q., Xu, J., Pan, Y., Chen, Q., and Wang, X. (2014). Hitsz-icrc at the ntcir-11 temporalia task. In *Proceedings of the NTCIR-11 Conference*.

[Jatowt and Au Yeung, 2011] Jatowt, A. and Au Yeung, C.-m. (2011). Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1259–1264, New York, NY, USA. ACM.

[Jatowt et al., 2009] Jatowt, A., Kanazawa, K., Oyama, S., and Tanaka, K. (2009). Supporting analysis of future-related information in news archives and the web. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 115–124, New York, NY, USA. ACM.

[Jatowt et al., 2010] Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., Kunieda, K., and Yamada, K. (2010). Analyzing collective view of future, time-referenced events on the web. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1123–1124, New York, NY, USA. ACM.

[Jensen and Snodgrass, 1999] Jensen, C. S. and Snodgrass, R. (1999). Temporal data management. *Knowledge and Data Engineering, IEEE Transactions on*, 11(1):36–44.

[Jin et al., 2008] Jin, P., Lian, J., Zhao, X., and Wan, S. (2008). Tise: A temporal search engine for web contents. In *Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium on*, volume 3, pages 220–224.

[Joho et al., 2014a] Joho, H., Jatowt, A., and Blanco, R. (2014a). Ntcir temporalia: a test collection for temporal information access research. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW)*, pages 845–850.

[Joho et al., 2014b] Joho, H., Jatowt, A., Blanco, R., Naka, H., and Yamamoto, S. (2014b). Overview of ntcir-11 temporal information access (temporalia) task. In *Proceedings of the NTCIR-11 Conference*.

[Jones and Diaz, 2007] Jones, R. and Diaz, F. (2007). Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3).

[Kahle, 1997] Kahle, B. (1997). Preserving the internet. *Scientific American*, 276(3):82–83.

[Kamps et al., 2004] Kamps, J., Marx, M., Mokken, R. J., and De Rijke, M. (2004). Using wordnet to measure semantic orientations of adjectives. In *LREC*, volume 4, pages 1115–1118. Citeseer.

[Kanazawa et al., 2011] Kanazawa, K., Jatowt, A., and Tanaka, K. (2011). Improving retrieval of future-related information in text collections. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 278–283, Washington, DC, USA. IEEE Computer Society.

[Kanhabua et al., 2011a] Kanhabua, N., Blanco, R., and Matthews, M. (2011a). Ranking related news predictions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 755–764, New York, NY, USA. ACM.

[Kanhabua et al., 2011b] Kanhabua, N., Blanco, R., and Matthews, M. (2011b). Ranking related news predictions. In *Proceedings of the 34th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 755–764.

[Kanhabua and Nørvåg, 2010] Kanhabua, N. and Nørvåg, K. (2010). Determining time of queries for re-ranking search results. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'10, pages 261–272, Berlin, Heidelberg. Springer-Verlag.

[Kanhabua and Nørvåg, 2012] Kanhabua, N. and Nørvåg, K. (2012). Learning to rank search results for time-sensitive queries. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2463–2466, New York, NY, USA. ACM.

[Kawai et al., 2010] Kawai, H., Jatowt, A., Tanaka, K., Kunieda, K., and Yamada, K. (2010). Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, pages 25:1–25:10, New York, NY, USA. ACM.

[Kim and Xing, 2013] Kim, G. and Xing, E. P. (2013). Time-sensitive web image ranking and retrieval via dynamic multi-task regression. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 163–172, New York, NY, USA. ACM.

[König et al., 2009] König, A. C., Gamon, M., and Wu, Q. (2009). Click-through prediction for news queries. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 347–354, New York, NY, USA. ACM.

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

[Leacock et al., 1998] Leacock, C., Miller, G., and Chodorow, M. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguisics*, 24(1):147–165.

[Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *5th Annual International Conference on Systems Documentation (SIGDOC)*, pages 24–26.

[Li and Croft, 2003] Li, X. and Croft, W. B. (2003). Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 469–475, New York, NY, USA. ACM.

[Liddy, 1998] Liddy, E. D. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science and Technology*, 24(4):14–16.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

[Luhn, 1957] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

[Luong et al., 2009] Luong, H. P., Gauch, S., and Speretta, M. (2009). Enriching concept descriptions in an amphibian ontology with vocabulary extracted from wordnet. In *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*, pages 1–6. IEEE.

[Maedche, 2002] Maedche, A. (2002). *Ontology learning for the semantic web*. Springer Science & Business Media.

[Mani et al., 2005] Mani, I., Pustejovsky, J., and Gaizauskas, R. (2005). *The language of time: a reader*, volume 126. Oxford University Press.

[Mani et al., 2004] Mani, I., Pustejovsky, J., and Sundheim, B. (2004). Introduction to the special issue on temporal information processing. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):1–10.

[Mani et al., 2001] Mani, I., Wilson, G., Ferro, L., and Sundheim, B. (2001). Guidelines for annotating temporal information. In *Proceedings of the first international conference on Human language technology research*, pages 1–3. Association for Computational Linguistics.

[Manning et al., 2008] Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

[Maron and Kuhns, 1960] Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

[Martin et al., 2014] Martin, P., Doucet, A., and Jurie, F. (2014). Dating color images with ordinal classification. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 447:447–447:450, New York, NY, USA. ACM.

[Matthews et al., 2010] Matthews, M., Tolchinsky, P., Blanco, R., Atserias, J., Mika, P., and Zaragoza, H. (2010). Searching through time in the new york times. In *Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval*, pages 41–44. Citeseer.

[McCarthy and Navigli, 2007] McCarthy, D. and Navigli, R. (2007). Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

[Metzler et al., 2009] Metzler, D., Jones, R., Peng, F., and Zhang, R. (2009). Improving search relevance for implicitly temporal queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 700–701. ACM.

[Miller, 1995] Miller, G. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Moens and Steedman, 1987] Moens, M. and Steedman, M. (1987). Temporal ontology in natural language. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 1–7.

[Mori et al., 2006] Mori, M., Miura, T., and Shioya, I. (2006). Topic detection and tracking for news web pages. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 338–342, Washington, DC, USA. IEEE Computer Society.

[Muller and Tannier, 2004] Muller, P. and Tannier, X. (2004). Annotating and measuring temporal relations in texts. In *Proceedings of the 20th international conference on Computational Linguistics*, page 50. Association for Computational Linguistics.

[Office, 1993] Office, U. S. A. R. P. A. S. . I. S. T. (1993). *Fifth Message Understanding Conference,(MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*. Morgan Kaufmann Publishers.

[Omelayenko, 2001] Omelayenko, B. (2001). Learning of ontologies for the web: the analysis of existent approaches. In *First International Workshop on Web Dynamics in Conjunction with the Eighth International Conference on Database Theory London, UK*, page 16.

[Orasan, 2006] Orasan, C. (2006). *Comparative evaluation of modular automatic summarisation systems using CAST*. PhD thesis, University of Wolverhampton.

[Palermo et al., 2012] Palermo, F., Hays, J., and Efros, A. A. (2012). Dating historical color images. In *Computer Vision–ECCV 2012*, pages 499–512. Springer.

[Pass et al., 2006] Pass, G., Chowdhury, A., and Torgeson, C. (2006). A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale)*.

[Perkiö et al., 2005] Perkiö, J., Buntine, W., and Tirri, H. (2005). A temporally adaptive content-based relevance ranking algorithm. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 647–648, New York, NY, USA. ACM.

[Plaisant et al., 1998] Plaisant, C., Shneiderman, B., and Mushlin, R. (1998). An information architecture to support the visualization of personal histories. *Information Processing & Management*, 34(5):581–597.

[Prager et al., 2006] Prager, J., Chu-Carroll, J., Brown, E. W., and Czuba, K. (2006). Question answering by predictive annotation. In *Advances in Open Domain Question Answering*, pages 307–347. Springer.

[Pustejovsky, 2005] Pustejovsky, J. (2005). Time and the semantic web. In *Temporal Representation and Reasoning, 2005. TIME 2005. 12th International Symposium on*, pages 5–8. IEEE.

[Pustejovsky et al., 2003] Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

[Pustejovsky et al., 2005a] Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., and Mani, I. (2005a). The specification language timeml. *The language of time: A reader*, pages 545–557.

[Pustejovsky et al., 2005b] Pustejovsky, J., Knippen, R., Littman, J., and Saurí, R. (2005b). Temporal and event information in natural language text. *Language resources and evaluation*, 39(2-3):123–164.

[Pustejovsky and Verhagen, 2009] Pustejovsky, J. and Verhagen, M. (2009). Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 112–116. Association for Computational Linguistics.

[Radinsky and Horvitz, 2013a] Radinsky, K. and Horvitz, E. (2013a). Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 255–264, New York, NY, USA. ACM.

[Radinsky and Horvitz, 2013b] Radinsky, K. and Horvitz, E. (2013b). Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM.

[Randolph, 2005] Randolph, J. (2005). Free-marginal multirater kappa (multirater κfree): an alternative to fleiss' fixed-marginal multirater kappa. *Joensuu Learning and Instruction Symposium*.

[Reiter and Buitelaar, 2008] Reiter, N. and Buitelaar, P. (2008). Lexical enrichment of a human anatomy ontology using wordnet. In *Proceedings of the Global WordNet Conference*.

[Robertson, 1977] Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304.

[Robertson et al., 1995] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109.

[Saha and Ekbal, 2013] Saha, S. and Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data Knowledge Engineering*, 85:15–39.

[Salton, 1971] Salton, G. (1971). The smart retrieval systemâĂŤexperiments in automatic document processing.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Schilder, 2004] Schilder, F. (2004). Extracting meaning from temporal nouns and temporal prepositions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):33–50.

[Schilder and Habel, 2001] Schilder, F. and Habel, C. (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, page 9. Association for Computational Linguistics.

[Schwarz and Morris, 2011] Schwarz, J. and Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1245–1254.

[Shah et al., 2014] Shah, A., Shah, D., and Majumber, P. (2014). Andd7 @ ntcir-11 temporal information access task. In *Proceedings of the NTCIR-11 Conference*.

[Shamsfard and Abdollahzadeh Barforoush, 2003] Shamsfard, M. and Abdollahzadeh Barforoush, A. (2003). The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review*, 18(04):293–316.

[Sheng et al., 2008] Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM.

[Shokouhi, 2011] Shokouhi, M. (2011). Detecting seasonal queries by time-series analysis. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1171–1172, New York, NY, USA. ACM.

[Siegel and Castellan, 1988] Siegel, N. and Castellan, J. (1988). *Nonparametric Statistics for the Social Sciences*. Mcgraw-hill edition.

[Singhal, 2001] Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

[Speretta and Gauch, 2008] Speretta, M. and Gauch, S. (2008). Using text mining to enrich the vocabulary of domain ontologies. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 1, pages 549–552.

[Strötgen et al., 2014] Strötgen, J., Armiti, A., Van Canh, T., Zell, J., and Gertz, M. (2014). Time for more languages: Temporal tagging of arabic, italian, spanish, and vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1.

[Strötgen and Gertz, 2010] Strötgen, J. and Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

[Strötgen and Gertz, 2012] Strötgen, J. and Gertz, M. (2012). Event-centric search and exploration in document collections. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 223–232. ACM.

[Strötgen and Gertz, 2013] Strötgen, J. and Gertz, M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation (LRE)*, 47:269–298.

[Svore et al., 2012] Svore, K. M., Teevan, J., Dumais, S. T., and Kulkarni, A. (2012). Creating temporally dynamic web search snippets. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1045–1046. ACM.

[Talukdar et al., 2012] Talukdar, P. P., Wijaya, D., and Mitchell, T. (2012). Coupled temporal scoping of relational facts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 73–82. ACM.

[Turtle and Croft, 1989] Turtle, H. and Croft, W. B. (1989). Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–24. ACM.

[UzZaman and Allen, 2010] UzZaman, N. and Allen, J. F. (2010). Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283. Association for Computational Linguistics.

[UzZaman et al., 2013] UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (* SEM 2013). Association for Computational Linguistics, June.*

[van Benthem, 1991] van Benthem, J. (1991). The logic of time, volume 156 of synthese library: Studies in epistemology, logic, methhodology, and philosophy of science.

[Verhagen, 2007] Verhagen, M. (2007). *Drawing TimeML Relations with TBox.* Springer.

[Verhagen et al., 2007] Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80.

[Verhagen et al., 2009] Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The tempeval challenge: Identifying temporal relations in text. *Language Resources and Evaluation (LRE)*, 43(2):161–179.

[Verhagen et al., 2005] Verhagen, M., Mani, I., Sauri, R., Knippen, R., Jang, S. B., Littman, J., Rumshisky, A., Phillips, J., and Pustejovsky, J. (2005). Automating temporal annotation with tarsqi. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 81–84. Association for Computational Linguistics.

[Yu et al., 2014] Yu, H.-T., Kang, X., and Ren, F. (2014). Tuta1 at the ntcir-11 temporalia task. In *Proceedings of the NTCIR-11 Conference.*

[Zhang et al., 2009] Zhang, R., Chang, Y., Zheng, Z., Metzler, D., and Nie, J.-y. (2009). Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 165–168, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zhang et al., 2010] Zhang, R., Konda, Y., Dong, A., Kolari, P., Chang, Y., and Zheng, Z. (2010). Learning recurrent event queries for web search. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1129–1139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Appendix A

Glossary

NLP	Natural Language Processing
IR	Information Retrieval
T-IR	Temporal Information Retrieval
PWN	Princeton WordNet
NIST	National Institute of Standards and Technology
TREC	Text Retrieval Conference
WWW	World Wide Web
PRP	Probabilistic Ranking Principle
MUC	Message Understanding Conference
NE	Named Entity
DTD	Document Type Definition
SGML	Standard Generalized Markup Language
TERQAS	Time and Event Recognition for Question Answering Systems
TANGO	TimeML Annotation Graphical Organizer
XML	Extensible Markup Language

TQIC	Temporal Query Intent Classification
STC	Sentence Temporal Classifier
GAs	Genetic Algorithms
SCE	Simple Classifier Ensemble
BVCE	Binary Vote based Classifier Ensemble
RVCE	Real/weighted Vote based Classifier Ensemble
IG	Information Gain
TempoWordnet Lexical	TWnL
TempoWordnet Probabilistic	TWnP
TempoWordnet Hybrid	TWnH
NTCIR	NII Test Collections for IR Systems
CLEF	Cross Language Evaluation Forum
TWn	TempoWordNet
IE	Information Extraction
QA	Question Answering

RÉSUMÉ DE LA THÈSE EN FRANÇAIS

La capacité à capturer l'information temporelle dans le langage naturel, qu'elle soit exprimée de manière explicite, implicite, ou par connotation, est essentielle pour de nombreuses applications telles l'extraction d'information, les systèmes de question-réponse, le résumé automatique. Associer une orientation temporelle au sens des mots pour capter l'information temporelle en langue est une tâche relativement directe pour les humains utilisant leurs connaissances sur le monde. Une base de connaissances lexicales associant automatiquement cette orientation au sens des mots serait de fait cruciale pour les tâches automatiques visant à interpréter la temporalité dans les textes.

Dans cette recherche, nous présentons une ontologie temporelle, TempoWordNet, où les synsets de WordNet sont enrichis avec une information sur leur temporalité intrinsèque : atemporel, passé, présent et futur. Nous étudions et expérimentons différentes stratégies de construction, lexico-sémantique, probabiliste et hybride.

TempoWordNet est évalué de maniére intrinsèque et extrinsèque, une ressource fiable devant à la fois contenir un étiquetage temporel de haute qualité et améliorer les performances de certaines tâches externes. Les deux types d'évaluations montrent la qualité et l'intérêt de la ressource. Pour compléter nos travaux, nous étudions aussi comment une application de recherche telle un moteur de recherche peut tirer parti de cette ressource.

Le retour des utilisateurs de TempoWordNet a encouragé à améliorer encore la ressource. Nous terminons donc en proposant une nouvelle stratégie de construction permettant d'améliorer de manière conséquente TempoWordNet.

MOTS-CLÉS

Traitement Automatique du Langage Naturel, Recherche de l'information, Bases de Données Spatio-Temporelles, Ontologies (Informatique)

RÉSUMÉ DE LA THÈSE EN ANGLAIS

The ability to capture the time information conveyed in natural language, where that information is expressed either explicitly, or implicitly, or connotative, is essential to many natural language processing applications such as information retrieval, question answering, automatic summarization, targeted marketing, loan repayment forecasting, and understanding economic patterns. Associating word senses with temporal orientation to grasp the temporal information in language is relatively straightforward task for humans by using world knowledge. With this in mind, a lexical temporal knowledge-base associating word senses automatically with their underlying temporal orientation would be crucial for the computational tasks aiming at interpretation of language of time in text.

In this research, we introduce a temporal ontology namely TempoWordNet where all the synsets of WordNet are augmented with their intrinsic temporal dimensions: atemporal, past, present, and future. We study and experiment different strategies to build TempoWordNet namely lexico-semantic, probabilistic, and hybrid. The resource is evaluated both intrinsically and extrinsically, the underlying idea being that a reliable resource must evidence high quality time-tagging as well as improved performance for some external tasks. Both the evaluations results confirm the quality and usefulness of the resource. To complement our research we also experiment how a search application can benefit from this resource.

Feedback from TempoWordNet users advocate for more reliable resource. At the end, we propose a strategy that shows steady improvements over the previous versions of TempoWordNet.

KEYWORDS

Natural Language Processing, Information Retrieval, Temporal Databases, Ontologies (Information Retrieval)

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen - GREYC
[UMR 6072]
Université de Caen Normandie
Campus Côte de Nacre, Boulevard du Maréchal Juin
CS 14032
14032 CAEN cedex 5
TEL : +33 (0)2 31 56 74 86
FAX : +33 (0)2 31 56 73 30