Parenthetical Classification for Information Extraction

Ismaïl El Maarouf, Jeanne Villaneau

Parenthetical Classification for Information Extraction

Ismail EL MAAROUF[1,2] *Jeanne VILLANEAU*[1]
(1) IRISA, Université Bretagne Sud, France
(2) RIILP, University of Wolverhampton, UK
elmaarouf.ismail@yahoo.fr, Jeanne.Villaneau@univ-ubs.fr

ABSTRACT
The article focuses on a rather unexplored topic in NLP: parenthetical classification. Parentheticals are defined as any text sequence between parentheses. They have been approached from isolated perspectives, like translation pairs extraction, but a full account of their syntactic and semantic properties is lacking. This article proposes a new comprehensive scheme drawn from corpus-based linguistic studies on French news. This research is part of a project investigating the structural aspects of punctuation signs and their usefulness for Information Extraction. Parenthetical classification is approached as a relation extraction problem split into three correlated subtasks: syntactic and semantic classification and head recognition. Corpus-based studies singled out 11 syntactic and 18 semantic relation subtypes. The article addresses automatic classification, using a combination of CRF and SVM. This baseline system reports 0.674 (head recognition), 0.908 (syntax), 0.734 (semantics), and 0.518 (end-to-end) of F1.

TITLE AND ABSTRACT IN ANOTHER LANGUAGE, FRENCH (FR)

Classification des parenthétiques pour l'extraction d'information

Définies dans cet article comme du texte entre parenthèses, les parenthétiques ont été jusqu'à présent peu étudiées en TALN. Si elles ont fait l'objet d'études particulières telles que l'extraction de paires de traduction, il manque une approche globale des relations syntaxiques et sémantiques qui les rattachent à leur contexte. Cet article propose un nouveau schéma de classification élaboré à partir d'études de corpus de presse. Cette recherche s'inscrit dans un projet explorant les aspects structurants des signes de ponctuation et leur utilité en Extraction d'Information. La classification des parenthétiques est abordée sous l'angle de l'extraction de relations et divisée en trois sous-tâches : classification syntaxique et sémantique et reconnaissances des têtes. Les études de corpus ont fait émerger 11 classes syntaxiques et 18 classes sémantiques. L'article propose d'évaluer un système combinant CRF et SVM. La baseline obtenue est de 0,674 (reconnaissance des têtes), 0,908 (syntaxe), 0,734 (sémantique) et 0,518 (toutes tâches confondues) de F-mesure.

KEYWORDS: Parentheticals, Punctuation, Information Extraction.

KEYWORDS IN FRENCH (FR): Parenthétiques, Ponctuation, Extraction d'information.

1 Condensed version in French (FR)

Cet article a pour objectif de contribuer à une meilleure connaissance des propriétés syntaxico-sémantiques des parenthétiques, définies comme des empans de texte entre parenthèses. La tâche y est abordée du point de vue de l'Extraction de Relations : (i) extraction des têtes externe et interne des parenthétiques, (ii) classifications syntaxique et sémantique des couples de têtes extraites.

La tête interne d'une parenthétique est son élément informationnel majeur. Sa tête externe est l'élément du contexte auquel l'information entre parenthèses doit préférentiellement être rattachée. Une particularité de ces têtes est de couvrir à peu près toutes les classes grammaticales : texte, phrase, Entités Nommées, noms, verbes, adjectifs, etc. Trois catégories spécifiques ont dû être définies pour les têtes externes : *a, p* renvoient respectivement aux cas où la tête externe est le texte entier et la phrase dans sa globalité alors que *n* renvoie au cas où il est impossible de spécifier un rattachement particulier. Dans les exemples de la Table 1, les têtes sont en caractères gras. La Table 2 (en bas à gauche) donne des précisions statistiques sur la nature des têtes dans le corpus étudié.

L'étude d'un corpus de la presse française (Le Monde) a permis d'identifier 11 classes syntaxiques, organisées suivant différents critères (voir détails section 4.2) : parenthétique de nature propositionnelle (inter-clause) ou non (intra-clause), apposition/adjonction (exemples (1, 4, 5, 8)/ (2, 3, 6, 7), présence ou absence de mots introductifs (soulignés dans les exemples), la parenthétique est (ou non) en coordination avec sa tête externe (exemples (3b), (7b)). La Table 1 donne un exemple de chacune des 10 principales classes ainsi définies.

	Inter-clause
(1)	le **produit intérieur brut** (**PIB**). [the **gross domestic product** (**GDP**).]
(2)	il est (**très**) **réussi**. [it is (**very**) **nice**.]
(3a)	son **taux** directeur (à 2,5%). [its reference **rate** (at 2.5%).]
(3b)	elle a connu la **liberté** (et les **pressions**). [She experienced **freedom** (and **pressure**).]
(4)	La cérémonie a **lieu** mercredi (cf. **page** 15) [Celebrations is **held** Wednesday (cf. **page** 15)]
	Intra-clause
(5)	elle est **partie** (Gustave **avait** 6 ans). [She **left** (Gustave **was** 6 years old).]
(6)	elle est **partie** ce jour-là (Gustave **ayant** 6 ans). [She **left** that day (Gustave **being** 6 years old).]
(7a)	elle est **partie** (alors que Gustave **avait** 6 ans). [She **left** (when Gustave **was** 6 years old).]
(7b)	elle est **partie** (et Gustave **avait** 6 ans). [She **left** (and Gustave **was** 6 years old).]
(8)	je ne **suis** pas (ici elle **baissa** la voix qui tremblait) de l'avis de sa majesté! [I **am** not (here she **lowered** her shaking voice) of your Highness's opinion!]

Table 1: Exemples pour la classification syntaxique. *(Examples for syntactic Classification.)*

La classification sémantique ne traite que d'une classe syntaxique particulièrement fréquente (82%) : les parenthétiques appositives non introduites et intra-propositionnelles (exemple (1) Table 1). Les études de corpus ont permis de mettre à jour 18 classes sémantiques génériques, comme l'ancrage spatial ou temporel (section 4.3).

Des conventions d'annotation ont été élaborées pour permettre l'annotation complète d'un corpus de 1000 parenthétiques. On pourra se reporter aux Tables 2 pour une description de ce corpus en termes de classes et à la Table 4 pour l'accord inter-annotateurs. Bien que, dans la classe des parenthétiques sémantiquement classées, la relation sémantique soit totalement implicite, le bon accord inter-annotateur montre, s'il en était besoin, que le lecteur décode sans difficultés la nature de l'information qui lui est donnée entre parenthèses.

Syntactic Class	Frequency
Intra App NI	801
Intra Adj IN Not-Coord	60
Inter App NI	27
Truncation	25
Intra Adj NI	21
Inter Adj IN NotCoord	22
Intra Adj IN Coord	21
Inter Adj NI	1
Total	978

Semantic Class	Frequency
NULL	177
CoRef-Abbreviation	150
Sit-SA	87
Cat-Instantiation	78
Sit-ArgVal	72
Sit-Affiliation	72
Ref-IR	55
CoRef-EntRef	49
Other	43
Sit-PS	43
Cat-ValPrec	28
CoRef-ValRef	27
Cat-Type	25
CoRef-Translation	22
Sit-TA-Date	21
Ref-PR	9
CoRef-Explanation	9
Sit-TA-Period	7
Ref-Coordinates	4
Total	978

Head Class	Frequency
ID	869
p	62
n	25
a	22
Total	978

Table 2: Fréquences des classes dans le corpus. *(Sample corpus class counts.)*

Le système proposé comme baseline (Section 6) combine les CRF (pour la détection des candidats) et les SVM (pour la classification), pour chaque tâche indépendamment et toutes tâches confondues. L'évaluation de ce système (Table 3) a permis tout d'abord d'observer que les ensembles de variables (formes, étiquettes morpho-syntaxiques, Entités Nommées, etc.) avaient un impact qui variait en fonction de la tâche : les étiquettes morpho-syntaxiques (T) sont par exemple les plus utiles à la classification syntaxique. De plus, la détection des candidats est une tâche cruciale, étant donné que le nombre de couples candidats aux frontières correctement délimitées est responsable d'une chute de la F-mesure globale du système (0,674, indépendamment, 0,518 toutes tâches confondues). Ces résultats sont confirmés par ceux obtenus par (Zhou et al., 2005) en Extraction de Relation à grand nombre de classes.

Feature	Pre-detection	Exact-Rec.	Soft-Rec	Syntax	Semantics
F	**0.965**	0.426	0.680	0.861	0.512
C	0.914	**0.499**	0.705	0.859	**0.637**
T	0.955	0.470	0.714	**0.908**	0.582
Ab	0.888	0.318	0.642	0.818	0.312
Pre-detection	-	0.349	**0.719**	-	-
Size	0.886	-	-	0.796	0.286
All	0.963	0.674	0.774	0.902	0.716
Baseline	0.888	0.3	0.649	0.818	0.182

Table 3: Résultats obtenus par le système en fonction des ensembles de variables utilisés. *(Independent task results on each feature set.)*

Les expériences à venir feront intervenir les informations syntaxiques pour évaluer leur apport. La robustesse du schéma d'annotation nécessite d'être également mise à l'épreuve; les expériences préliminaires menées en ce sens sur des corpus encyclopédiques, littéraires, juridiques et scientifiques n'ont jusqu'à présent donné lieu qu'à des modifications minimes.

2 Introduction

As Say and Akman (1996) point out, punctuation has not attracted much theoretical attention in Linguistics nor in NLP (see however (Briscoe, 1996; Jones, 1996; Nunberg, 1990)). Nevertheless, it is pervasive in written texts and commonly used in NLP systems: phrase boundaries, sentence boundaries, and so on.

This work is a part of a discourse-oriented project investigating how punctuation interacts with different linguistic levels such as syntax and semantics. It attempts to provide answers as to why and how punctuation helps comprehension, through the analysis of text segments between parentheses, named parentheticals. This punctuation structure should not be confused with the definition of parentheticals as optional embedded segments.

The article introduces a new scheme designed for robustness and large coverage from an Information Extraction perspective. The task is divided into three subtasks, Head recognition, Syntactic and Semantic classification. The choice of the classes is based on a linguistic corpus study on French news: 11 syntactic relations and 18 semantic relations have been defined, according to several levels of granularity or dimensions.

The article describes the application and results of an annotation experiment on a sampled corpus of a thousand parenthetical observations. It provides baselines for each subtask, using various feature sets. The results, along with an analysis of feature set impact, call for further experiments as well as for a generalization of the task to different corpora.

Section 3 discusses related work on parentheticals and section 4 introduces the classification scheme. Section 5 details the annotated corpus. The systems are described in section 6 and their evaluation is presented in section 6.3.

3 Parentheticals in Information Extraction (IE)

It is commonly stated that parentheticals provide optional information: they can be removed without affecting understanding. For instance, they are deleted in sentence compression applications (e.g. equation (31) in (Clarke and Lapata, 2008)). However, they have recently aroused interest in IE.

IE (Sundheim, 1991; Sarawagi, 2008) is the NLP field concerned with (i) the identification of Named Entities (NE) from text, (ii) their co-referring units (anaphora, acronyms), and (iii) their interactions (e.g. Affiliation, Location). Two text types have particularly been studied in IE: newswire and biomedical articles. In both types, parentheticals are pervasive. For instance, Bretonnel Cohen et al. (2010), report finding about 17,000 parentheses in a corpus of 97 scientific articles (about 600,000 words). Comparatively, we found more than 4 parentheses per article in newswire texts (136,000 on 17,000,000 words).

Parentheticals have mainly been studied under the topics of Abbreviation, Translation and Transliteration pairs extraction. Abbreviation recognition (extracting co-referring full and short forms) is a well-defined task which has both been conducted on biomedical literature (Pustejovsky et al., 2001; Schwartz and Hearst, 2003) and on newswire (Okazaki et al., 2008): systems generally record more than 0.9 in F1. Okazaki et al. (2008) analyzed 7,887 frequent parenthetical instances and classified them into *Acronym*, *Translated Acronym*, *Alias* and *Other*. In their study, the *Other* category covers 81.9% of all studied instances. The authors propose to split it into *alphabetic transcription*, *location*, or *affiliation*.

Parentheticals have also been studied in the field of Machine Translation. Cao et al. (2007)

observe that many terms (very frequently NE) are followed by their English translation inside parentheses on Chinese monolingual webpages. They use parentheticals to extract a bilingual dictionary automatically, and find that the majority of pairs are not covered by a standard lexicon. In a similar experiment, Kaji and Kitsuregawa (2011) propose to classify transliteration pairs in order to help segmenting complex katakana compounds.

A recent much larger scheme was proposed for the biomedical domain (Bretonnel Cohen et al., 2011). The authors propose to classify parenthetic content into 20 categories. They note that some categories are ambiguous if only the content inside parentheses is taken into account. The scheme introduced in the next section builds on previous works and aims to be generic and rich at the same time.

4 Annotation scheme

The annotation scheme is the result of an in-depth corpus-based linguistic study. It proposes to identify, when possible, the most prominent unit inside parentheses (internal head) and the word in the host sentence (external head) to which it is most preferably linked. This link is syntactically characterized. Besides, most of the time, deleting parentheses affects sentence grammaticality (cf. ex. (4)), so the relation between parenthetical and its environment needs to be inferred by the reader. In this case (and only in this case), the scheme provides semantic categories.

4.1 Head Detection

Internal heads are most straightforwardly detected, because they tend to correspond to the syntactic head of the first information group. When more than one head can be selected, only the first is kept.

External heads are very frequently multi-word units (cf. example (1), Table 1), but is not necessarily the head of its own syntactic phrase. In the following example (9), the relation holds between a color and its interpretation, but it is "niveau" which is the syntactic head of the prepositional phrase.

(9)...maintenir le niveau d'alerte antiterroriste au niveau **orange** *(très* **élevé***) [keeping the antiterrorism threat level at level* **orange** *(very* **high***)]*

In some rarer cases, the external head may follow (and not precede) the parenthetical, as in ex. (2), Table 1. More examples can be found in bold type in Table 1.

Three labels are provided when no words can be singled out as head: *p* for the whole proposition, *n* for no head, for example in the case of truncation (cf. end of 4.2) and *a* is used when the parenthetical provides information on the whole document.

4.2 Syntactic classification

Ten syntactic categories were organized along four criteria. An example for each of them is given Table 1.

- The first criterion is the distinction between *intra*(-clause) and *inter*(-clause). A parenthetical is *inter* if its content can be viewed as a finite clause (cf. examples (5), (6), (7a), (7b) and (8) of Table 1.). In contrast, an *intra* corresponds to non-finite clauses (cf. (1), (2), (3a), (3b) and (4)).

- The second criterion discriminates between *adj*(-oined) and *app*(-ositionnal) (non-adjoined) parentheticals. In the case of *adj* parentheticals, the sentence remains correct when the brackets are removed (cf. (2), (3a), (3b), (6), (7a) and (7b), Table 1.). In *app* parentheticals, the deletion of the parentheses breaks the progression of the sentence (cf. (1), (4), (5) and (8)).
- The third criterion divides parentheticals into *intro*(-duced) (IN) and *not-intro*(duced) (NI) parentheticals. A parenthetical is *intro* when an expression introduces its head, and links it with the outer context (cf. (3a), (3b), (4), (7a), (7b) and (8) of Table 1, where introducing elements are underlined).

Eight classes are obtained by applying the previous three criteria. A fourth criterion splits *intro adj* parentheticals (*inter* or *intra*) and discriminates between *coord*(-inated) and *not-coord*(-inated) parentheticals. In *coord* parentheticals, the internal head has the same syntactic category as the external head (word or clause) (cf. (3b) and (7b)).

The last and eleventh class concerns the case of punctuation marks in brackets ((...), (!), etc.), called *truncation*. All cases have been found in corpus, though with high distribution differences (Table 2, left).

4.3 Semantic classification

Eighteen semantic categories, organized into four dimensions, were defined for *intra app NI (intra-clause appositional not-introduced)* parentheticals, which lack an explicit link. Classifying other syntactic classes was left for further investigation.

1. The first, *Co-reference (CoRef)*, corresponds to cases where both heads refer to the same entity, but use different names.
   (a) *Abbreviation*: the parenthetical contains an abbreviation of the external head (its full form; cf. example (1)).
   (b) *Explanation*: the definition of an acronym (the reverse of the previous relation).
   (c) *Translation*: it contains a translation of the external head in an other language.
   (d) *Reformulated Entity (RefEnt)*: other co-referential relations not covered by the previous classes; for example, the name an actor has in a movie.
   (e) *Reformulated Value (RefVal)*: it translates the value expressed by the external head in another unit of measurement.
2. The second broad class, *Categorization (Cat)*, refers to asymmetric relations between entities and categories.
   (a) *Type*: it provides the category of the entity of the external head (as hyponyms).
   (b) *Instantiation*: the reverse of the previous relation. It provides an instance of the category expressed by the external head.
   (c) *Value Precision (ValPrec)*: it precises the value of its external head, which is already a quantity category (drop, growth, etc.).
3. The third class relates to *Situational relations (Sit)*. Most correspond to standard semantic relations defined for relation extraction (ACE, 2008).
   (a) *Product Source (PS)*: it refers to the producer, editor, etc. of a product referred by the external head (e.g. book).

(b) *Affiliation*: it contains the organization to which its external head (person or organization) is affiliated.
(c) *Spatial Anchoring (SA)*: it sets the spatial location of an entity.
(d) *Temporal Anchoring (TA)* is split into *Date* and *Period* (of any kind of entity).
(e) *Argument Value (ArgVal)*: it gives a value related to its external head (as age).

4. The fourth class concerns *Referencing (Ref)*, where parentheticals attribute references or indexes to the external head.

(a) *Inter-textual Reference (IR)*: it makes a reference to the journal, media as source of the external head (citation).
(b) *Para-textual Reference (PR)*: it refers to para-textual elements of the document (figure, footnote, etc.)
(c) *Coordinates*: It provides the code value indexing entities in a given coding scheme (phone number, postal address, etc.).
(d) *Indexing*: it refers to the marks (numbers) indexing document elements (such as examples) and to which parentheticals may elsewhere refer to.

Contrary to Okazaki et al. (2008), translated acronyms are here considered as abbreviations. In principle, most classes defined by Bretonnel Cohen et al. (2011) could be fitted in this scheme, like *p-values* (*ArgVal*) or *Figure references* (*IR*).

5 Corpus Annotation

The scheme was tested on a sample of French news (114 parentheticals) by two highly-trained annotators. The results of inter-annotator agreement for the three tasks are illustrated in Table 4. Kappa indexes show that parenthetical syntactic (0.89) and semantic (0.79) categories could easily be recognized by annotators. The Kappa was not computed for Head recognition since head spans vary greatly. It is thus hard to approximate the random baseline on which the Kappa is based (Grouin et al., 2011).

Task	# agr.	# disagr.	Total	Kappa
Syntax	109	5	114	0.89
Semantics	88	13	101	0.79
Head	103	11	114	/

Table 4: Inter-annotator agreement synthesis.

As can be seen in Table 2 (left), the *intra app NI* class is the most frequent syntactic class. This validates the use of a semantic scheme designed especially for this class (other syntactic classes being semantically classified as *NULL*). Heads are mostly words, though the "p" class covers 6% of examples.

The counts of semantic classes (Table 2, right) shows that the semantic class *Other*, used for the examples of *intra app NI* parentheticals which don't match with the defined semantic categories covers less than 5% of examples.

At last, the annotated corpus was sampled (stratified sampling) according to the concatenated labels to build the training and testing corpora (half each).

6 System design and Evaluation

6.1 Overview

Relation Extraction (RE) systems typically (i) extract Named Entity (NE) pairs to filter positive targeted instances (recognition step), before (ii) they attribute a label to them (classification step). The recognition step is problematic since it requires that all possible NE instances be extracted: Sun et al. (2011) indicate that the number of negative instances is about 8 times higher than the number of positive ones. The current best classification systems on complex schemes rely on feature-based approaches (Zhou et al., 2005). Such methods typically use information on candidate NE pairs (such as NE tag, POS tag, form, etc.), along with information on the words in between (Zhou et al., 2005) for prediction.

In our case, candidate pairs (heads) do not correspond uniquely to NE, but also to whole sentences, quotations, verbs, adjectives, etc.: the number of candidate pairs is huge. This is why, instead of elaborating a preprocessing system, the recognition step was approached as a sequence labeling task (6.2).

What is more, annotators had the choice between using labels and select word spans to identify parenthetical heads. Therefore, the system first discriminates labeled instances (*a*, *p*, and *n*) from others (*ID* class). In a second step, it detects head boundaries from previously pre-detected *ID* instances. This first step (pre-detection), along with syntactic and semantic classification, is approached as a classification task performed on each parenthetical instance.

6.2 System

Two systems were used : CRF++ (Kudo, 2007) for head recognition and SVM (Hall et al., 2009) for parenthetical classification as they are recognized as very efficient algorithms [1]. The features used for CRF Recognition include:

- forms (F) without any processing.
- categories (C) provided by a linguistic analyzer, which includes NE recognition and semantic labels (Rosset et al., 2006). This tagset was transformed into BIO format (Tjong Kim Sang and De Meulder, 2003).
- POS tags (T) provided by the Tree-tagger (Schmid, 2003).
- Abbreviation pairs (Ab) from the system provided by Schwartz and Hearst (2003).
- pre-detection labels (*a*, *p*, *n*, *ID*) propagated on all the words,

Unigrams, bigrams, and label bigrams (Kudo, 2007) occurring in the most optimal window size (cf. 6.3.2) were used for all feature sets.

The same features were used for classification, except removing predetection labels and adding parenthetical size (*Size*). For the other sets (*F*, *C*, and *T*), each feature value was combined with positional parameters to distinguish between the first and second words before and after the opening brace.

6.3 Evaluation

Evaluation was performed on the test corpus (490 instances) using the standard metrics of precision, recall and F1 (F-measure). All results are displayed in Table 3.

[1] Different algorithms were tested to confirm this.

6.3.1 Head Label pre-detection

Pre-detection is a straightforward task: most corpus instances are annotated with one label (ID), which results in a high baseline of 0.888, just by assigning this label to all examples. The SVM beats this baseline with 0.963 of overall F1. Detailed feature analysis shows that the *Ab* and *Size* features do not individually help for this task since the resulting models behave like the baseline. The best feature set is F (forms): the SVM perfectly classifies *a* and *n* classes (1 in F-measure). This is due to the fact that the *a* class corresponds to a parenthetical which only contains punctuation signs such as "...". The *n* class instances generally occur at the end of an article and are immediately preceded by dashes. The real challenge is therefore to discriminate the *p* class (0.667) from the *ID* class (0.981).

6.3.2 Head recognition

Only external heads were evaluated for this task. The baseline selects the word immediately preceding the parenthetical as the head, because most heads occupy this position. An example can be considered correctly labeled if (i) all the labeled words need to be correct (exact evaluation), or if (ii) at least one word needs to be correctly labeled (soft evaluation). The baseline F1 is very low (0.3) in the first case, and reaches 0.649 in the second case. The best results (0.674) recorded for the CRF were obtained with a window size of 4 words [-1,2]. The best feature set is *C* (0.499), i.e. the categories provided by the linguistic analyzer, including Named Entities. These results are still much lower than the system using the combination of all feature sets (0.674 in F1 for exact matching; 0.774 for soft matching). The latter takes benefit of the pre-detection features (best feature set for *p* and *n* classes) but also largely improves exact head recognition (+0.146 compared to *C*).

The high difference between Soft and Exact head recognition across feature sets indicates that multi-word units management play a large part in system performances.

6.3.3 Classification

The Syntax and semantic tasks were first carried out independently. The Syntax task consists of 7 labels (4 rare *inter* categories are missing) and the semantic task, of 19 classes (*Indexing* is missing). The baseline model assigns the most frequent class to all examples (0.818 in F1 for syntax, 0.182 for semantics). Table 3 shows the superiority of the *T* set for syntax. *T* is composed of precise syntactic labels; for instance, it discriminates between various verb forms such as past and present participles (contrary to the *C* set which only divides between auxiliaries, modals, actions and gerunds).

Concerning Semantics, it is the *C* feature set which is the most effective. This said, the system reaches higher scores when all the features are taken together. It is also clear from the table that POS tags (*T*) have a greater impact than forms (*F*) on this task.

A second experiment was conducted to analyze the impact of syntax on semantics: only the examples predicted as *Intra App NI* (the most frequent class to be semantically labeled) by SVM-T were extracted for semantic classification (the rest being considered as *NULL*). This filtering method prove successful (0.734; +0.018 improvement): even if 8 examples are incorrectly filtered (semantically *NULL*), the system correctly classifies 31 semantic instances. Detailed class analysis indicate that improvements mostly affect *ValPrec* (+0.22), *NULL* (+0.2), and *Other* (+0.15).

6.3.4 End-to-end Evaluation

The aim of the end-to-end evaluation is to observe how Head recognition affects both syntactic and semantic classification. An example was considered correct when all task labels were correctly assigned. F1 significantly drops drops to 0.518 on exact matching, and to 0.586 on soft matching. These results are consistent with previous work in RE. Zhou et al. (2005) report 0.55 of F1 when recognition and classification are evaluated together on subtypes (0.68 on supertypes), and attribute 73% of errors to recognition (53% in our case).

It is interesting that the *Situational* dimension, which contains traditional RE broad categories (*SA* and *Affiliation*), obtains the best scores. These scores are even higher than reported in RE literature (Sun et al., 2011), though the dataset is barely comparable. *Abbreviation* experiences comparably lower results than reported in the literature: Okazaki et al. (2008) report 95.7% accuracy (0.887 of F1).

7 Conclusion and discussion

Parenthetical classification is a rather unexplored topic and this article aims at providing insights into this punctuation pattern. An annotation scheme was designed to cover most frequent cases for three tasks: syntactic and semantic classification and head recognition.

Corpus analyses revealed that most parentheticals lack an explicit link to the external context (the *App* syntactic class), but are nonetheless similarly understood by annotators. Only the *Intra App NI* class was semantically labeled (81% of instances) and tested. Analyzing *inter app* parentheticals was left for further investigation because it is believed that they must be studied on the discourse level (see for example (Marcu, 2000)): proposition links may be characterized as *causal* for instance.

Other annotation experiments have been started on different text types (encyclopedic, legal, scientific or fictional documents), to assess the robustness of the scheme across text types, and evaluate automatic systems in the light of domain adaptation. Preliminary results are encouraging in the sense that the same scheme can be used with little adaptation.

The evaluation proposed a baseline using CRF and SVM for each task separately, with various feature sets based on POS tags, Named Entities, Forms, etc. The best model reported 0.908 for syntax, 0.734 for semantics, and 0.674 for head recognition. It is interesting that different feature sets have had different impacts on classification tasks. All tasks except semantics have shown better performance on isolated feature sets. Besides, Zhou et al. (2005) have shown that chunking improves performances ACE Relation Extraction. Following evaluations should investigate the benefits of feature sets like chunking and semantic lexicons (as hyperonym lexicon for *Type* and *Instanciation* categories).

Since classification tasks such as syntax or semantics reported better results, it would also be interesting to investigate what gain results from their use as feature sets, much like what was done for pre-detection. Overall, it seems that improving recognition performances would rely on careful feature construction.

As suggested in section 6.3.4, the results obtained for *Affiliation* and *SA* are higher than usually reported on standard RE. This could simply be due to the fact that parenthetical structures impose strong constraints which facilitate classification. If these results are confirmed in subsequent evaluations, it would mean that parentheticals could be used as a small window to extract valuable seeds for general RE.

References

ACE (2008). Automatic content extraction 2008 evaluation plan. assessment of detection and recognition of entities and relations within and across documents.

Bretonnel Cohen, K., Christiansen, T., and Hunter, L. E. (2011). Parenthetically speaking: Classifying the contents of parentheses for text mining. In *AMIA annual symposium proceedings: 267*.

Bretonnel Cohen, K., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics, 11:492*.

Briscoe, T. (1996). The syntax and semantics of punctuation and its use in interpretation. In *Proceedings of the ACL Workshop on Punctuation:1–7*.

Cao, G., Gao, J., and Nie, J.-Y. (2007). A system to mine large-scale bilingual dictionaries from monolingual web pages. In *MT summit XI proceedings: 57-64*.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research (JAIR), 31:399–429*.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR. Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations, 11(1)*. `http://www.cs.waikato.ac.nz/ml/weka`.

Jones, B. (1996). *What's The Point? A (Computational) Theory of Punctuation*. PhD thesis, University of Pennsylvania.

Kaji, N. and Kitsuregawa, M. (2011). Splitting noun compounds via monolingual and bilingual paraphrasing: A study on japanese katakana words. In *EMNLP proceedings: 959–969*.

Kudo, T. (2007). Crf++: Yet another crf toolkit. `http://crfpp.sourceforge.net`.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics, 26(3):395–448*.

Nunberg, G. (1990). *The Linguistics of Punctuation*. CSLI Lecture notes (18), CSLI publications, Stanford, CA.

Okazaki, N., Ishizuka, M., and Tsujii, J. (2008). A discriminative approach to japanese abbreviation extraction. In *IJCNLP proceedings: 889–894*.

Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M., and Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from medline databases. *Studies in health technology and informatics, 84:371–375*.

Rosset, S., Galibert, O., Illouz, G., and Max, A. (2006). Integrating spoken dialog and question answering: the ritel project. In *Proceedings of InterSpeech'06*.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases, 1(3):261–377*.

Say, B. and Akman, V. (1996). Current approaches to punctuation in computational linguistics. *Computers and the Humanities, 30(6):457–469*.

Schmid, H. (2003). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP 1994 proceedings:44–49*.

Schwartz, A. S. and Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing 84:451–462*.

Sun, A., Grishman, R., and Sekine, S. (2011). Semi-supervised relation extraction with large-scale word clustering. In *ACL 2011 proceedings: 521–529*.

Sundheim, B. M. (1991). Overview of the third message understanding evaluation and conference. In *Proceedings of MUC:3–16*.

Tjong Kim Sang, E. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *CoNLL 2003 proceedings: 142–147*.

Zhou, G., Su, J., Zhang, J., and Zhang, M. (2005). Exploring various knowledge in relation extraction. In *ACL 2005 proceedings*.