Improving MT coherence through text- level processing of input texts: the COMTIS project. Session 6 - Translation and Natural Language Processing

Andrei Popescu-Belis, Bruno Cartoni, Andrea Gesmundo, James Henderson, Cristina Grisot, Paola Merlo, Thomas Meyer, Jacques Moeschler, Sandrine Zufferey

TRALOGY

Improving MT coherence through text-level processing of input texts: the COMTIS project

Andrei Popescu-Belis
Idiap Research Institute, Martigny, Switzerland

Bruno Cartoni
Department of Linguistics, University of Geneva, Switzerland

Andrea Gesmundo
Department of Computer Science, University of Geneva, Switzerland

James Henderson
Department of Computer Science, University of Geneva, Switzerland

Cristina Grisot
Department of Linguistics, University of Geneva, Switzerland

Paola Merlo
Department of Computer Science, University of Geneva, Switzerland

Thomas Meyer
Idiap Research Institute, Martigny, Switzerland

Jacques Moeschler
Department of Linguistics, University of Geneva, Switzerland

Sandrine Zufferey
Department of Linguistics, University of Geneva, Switzerland

TRALOGY I - Session 6
Date d'intervention : 04/03/2011

This paper presents an ongoing research project, started in March 2010 and sponsored by the Swiss National Science Foundation, which aims at improving machine translation output in terms of textual coherence. Coherence in text is mainly due to inter-sentential dependencies. Statistical Machine Translation (SMT) systems, currently sentence-based, often fail to translate these dependencies correctly. Within the COMTIS project, state-of-the-art linguistics research and Natural Language Processing (NLP) techniques are combined to identify and to label inter-sentential dependencies that can be learned by SMT system in the training phase.

lien video : https://webcast.in2p3.fr/video/the_comtis_project

Introduction

Machine translation has made significant progress in the past decade, but its focus has remained on the translation of sentences considered individually. However, one of the key-features of a multi-sentence text's quality is its *coherence*. From a linguistic point of view, textual coherence is determined by a certain number of aspects, such as pronouns and other referring expressions, verb tense/mode/aspect, discourse relations which may be signalled by connectives, and politeness/style/register. All these aspects have to be addressed in an inter-sentential perspective, modelling the internal coherence of the text for translation.

The article is structured as follows. In the first section, we provide a brief overview of the state-of-the-art that motivates the COMTIS project. In the second section, we describe the COMTIS project and its different modules that are combined to improve the translation of cohesion aspects in text. In the third section, we present some of the first experiments with annotation of English connectives, which set the methodological grounds for the other aspects under consideration in COMTIS.

1. Motivation for the COMTIS project

In NLP, a fair number of studies address *coherence* both in automatic language analysis and generation. For the purpose of language analysis, as a pre-processing step, discourse structure was hand-annotated in the Penn Discourse Treebank (Prasad et al., 2008), based on explicit discourse connectives and implicit discourse relations. Research–on cohesion ties has focused on specific phenomena such as anaphora, Mitkov (2002); lexical chains (Morris and Hirst 1991); temporal relations (Lapata and Lascarides 2004); discourse markers (Litman 1996), and more specifically on the disambiguation of connectives (Hutchinson 2004). Cohesion has probably gained more attention from a language generation perspective, mainly because it is a characteristic that must be explicitly managed when generating discourse. Therefore, any sentence planner system must somehow make decisions on cohesion markers (Hovy, 1988). Methods for correctly selecting tense as well as temporal connectives have been proposed for English by Dorr and Gaasterland (1995) and for German by Grote and Stede (1998).

For machine translation of inter-sentential phenomena, fewer attempts have been made, especially in situations when ignoring such phenomena was very detrimental to MT quality, for rule-based systems. For instance, Japanese zero pronouns must be overtly expressed when translating Japanese texts into English, and experiments with the integration of anaphora resolution within a transfer-based MT system have been made for extra-sentential antecedents (Nakaiwa and Ikehara, 1992), intra-sentential ones (Nakaiwa and Ikehara, 1995), and for deictic pronouns (Nakaiwa and Shirai,1996). Experiments have also been made with anaphora in an interlingua-based Spanish to English translation system (Peral et al., 1999). A study of intrasentential disambiguation of the referential properties of nouns for Japanese to English MT has been done by Murata and Nagao (1993) and Murata et al. (2001). Translation of aspectual information (state vs. event, perfective vs. progressive) has received knowledge-based modelling in relation to lexical-semantic information (Dorr, 1992), in Dorr's Unitran English/Spanish interlingua-based MT system (Dorr, 1993). However, in this approach, aspect was processed on a purely sentence-by-sentence basis. Marcu et al. (2000) have proposed a model for the translation of discourse structure, in RSTstyle (Mann and Thompson, 1988), with the goal of rewriting discourse structures from Japanese texts into discourse structures that are more natural to English. The model was implemented as a module intended for a discourse-based MT system, which should have been accompanied by a Japanese discourse parser and a statistical translation module combining discourse-specific features and standard SMT models.

In the COMTIS project, the main objective is to combine inter-sentential information with statistical machine translation models. We analyze the discourse-level phenomena that influence

the perceived coherence of a text, and we plan to use surface cues to label them automatically with inter-sentential dependency labels (ISD). The labelled source text will then be used for training new SMT models which are capable to learn the ISDs, and then when translating a new text. Further description of all the steps of the project is provided in the following section.

2. Description of the COMTIS project: aims and organisation

The COMTIS project addresses the discourse-level phenomena that determine the coherence of texts and which at the same time are difficult to translate, i.e. verbal tense/aspect/mode, discourse connectives and pronouns. Once identified and taxonomized, the inter-sentential dependency (ISD) phenomena are labelled in text: first manually, for use in development and training of automatic labellers, and then using machine-learning based classifiers. SMT systems are also trained to work with such labelled texts. To translate a new text, it is automatically labelled with ISDs and then handed over .to an SMT system that can process the ISD labels, in order to obtain a more coherent translation. Evaluating issues are also addressed.

The COMTIS project focuses mainly on the English-French language pair (in both translation directions) although other languages will also be considered, namely German and Italian. The following objectives, described in the next subsections, will be pursued in the project. A global picture of the research plan is given at the end of the section. At the current stage of the project, only connectives and verbal tense/aspect/mode have been studied. Examples will focus therefore on these two aspects.

2.1 Topic 1: Linguistic analysis of inter-sentential relations

Only connectives and verbal tense/aspect/mode issues have been addressed so far in COMTIS. Problematic connectives (i.e. connectives that convey more than one "discourse relation" and that do not have a unique equivalent in the target language) have been identified and extended corpus research has been performed to shed new light on their different meanings and contexts of usage. Section 3 below provides further insights on the precise ongoing pieces of research.

For the verbal tense/aspect/mode issues, research focuses on past and present tenses that appear to display the most problematic divergences for each direction of translation, from English to French and from French to English. For example, one of the divergences is that the translations of the English Simple Past do not equivocally match the French verbal system (which has "passé compose", "imparfait", and "passé simple"), as shown in examples 1a, 1b (The *JRC-Acquis Multilingual Parallel Corpus*) and 1c (Wilde, "The portrait of Mr. W.H.", page 24) with English as source language.

1a. "I wish to express that view even if I respectfully disagreed and voted against the proposal of the President of the Socialist Group."	*Je souhaite exprimer ce point de vue, même si je désapprouve la proposition du président du groupe des socialistes, tout en la respectant, et même si j'ai voté contre.*
1b. "A lot of them have not signed up to or not ratified the Convention on the Protection of Financial Interests, and therefore it was clear that something more radical needed to be done."	*De nombreux États membres n' ont pas ratifié la convention sur la protection des intérêts financiers ou n' ont pris aucun engagement à cet égard et , dès lors , il était clair qu' il fallait entreprendre quelque chose de plus radical.*

1c. "Erskine remained silent for a few moments looking at the thin grey threads of smoke that were rising from his cigarette."	*Erskine garda le silence quelques instants, observant les fines volutes de fumée grise qui s'élevaient de sa cigarette.*

Other divergences are also considered, such as the translation of the French imperfect (by English simple past or past continuous) or the translation of English continuous tenses in French.

The analysis of both research problems above (connectives and verb tenses) are based on large literature reviews of previous research in linguistics. For instance, verb tenses are described and analyzed in terms of aspectual and temporal properties, such as Vendler's classification of verbs in 1967 (lexical aspect), grammatical aspect (*perfect* and *imperfect/progressive*) and Reichenbach's temporal coordinates *point of speech, point of reference* and *point of event* (1947). Tense are seen as a referential category and are interpreted in the Relevance Theory framework (Sperber & Wilson, 1986), in that they direct the hearer in the process of recognising the speaker's communicative intention. We will get back to connectives in Section 3 below.

Large corpus-based contrastive analyses are also performed for the different phenomena, in order to highlight main divergences, which, when solved, would have an impact on translation quality.

2.2 Topic 2: annotation of corpus data

Within this topic, the identified phenomena in Topic 1 are addressed by corpus-based analysis. This task mainly involves gathering enough annotated data for the following topics. In doing so, this provides information about the importance of the considered phenomena (only "frequent" ones should be addressed), and highlight interesting new knowledge not caught by the theoretical studies in Topic 1, but revealed by corpus data.

For discourse connectives, different methods of manual annotations have been tested, and are described in Section 3. They rely on existing descriptive resources and their own annotation manuals, such as the PDTB (Prasad et al., 2008) for English or the LexCONN database (Roze *et. al 2010*) for French. Particular focus is also put on the influence of the translation direction, and consequently on the status of the text (original or translated), aspects that influences the kind of connectives found in context.

For the tense/aspect/mode issue, extensive contrastive research is currently being carried out to highlight precise translation issues that need to be addressed, and to test existing annotation and formalisation frameworks (such as Reichenbach "point of reference", and TimeML annotation schemes) to see if they can convey disambiguation information and can appropriately be used as inter-sentential dependency labels. Different registers are also studied, as not every discourse register and style makes use of the same kind of verbal relation.

2.3 Topic 3: automatic identification of inter-sentential dependencies

Based on the annotations performed in Topic 2, a disambiguation module will be built in order to automatically identify and tag ISDs in texts. The task consists of designing, training and evaluating statistical and algorithmic classifiers based on robust surface features for cohesive markers: anaphora, verbal tense/mood/aspect, discourse connectives, and register/style. The methods that will be used at this level will be inspired by the state-of-the art (e.g., for connectives: Lin et al. 2010; Pitler and Nenkova 2009, for time/aspect: Dorr and Gasterland 1995), with the goal of implementing robust, operational classifiers that are principally useful for labelling cohesion markers in preparation for MT.

The results of the disambiguation module are 'explicitation' tags (e.g. inspired by the PDTB annotation or the TimeML standard) and will be adjoined to the ISDs in texts and will be used by the statistical MT engine. These tags will disambiguate those cohesion markers that are potentially relevant to translation, as identified by the initial analyses under Topic 2, using also the features indicated by the linguistically-motivated approach in Topic 1 and features from synchronous parsing performed in Topic 4.

The possibility of a joint classification of the four types of inter-sentential phenomena should not only increase the classification accuracy but also provide the important empirical confirmation of the relevance of the overall COMTIS project hypotheses.

2.4 Topic 4: Statistical machine translation for ISD-labelled texts

Once enriched with ISD labels, disambiguated text is processed by an SMT system that has been trained on such kind of texts (produced in Topic 2 and 3).

Different pre-processing approaches will be considered for this. First, an SMT system not exploiting the ISD annotations will be built as a baseline in order to be able to directly compare the translation quality gain from ISD-labelled texts.

On the way to a system capable of processing ISD-labelled texts, there are several methods to be evaluated.

A first method will simply enter the ISD-labelled connectives into an existing phrase table (marking the words and adjusting the probabilistic weights in the phrase table of an already trained SMT decoder). A second way is to tag a large parallel and aligned corpus with the discourse information obtained from the disambiguation module and then to train a new SMT system learning and weighting these tags during training. The latter has been done for pronouns by Le Nagard and Koehn (2010) or for the reordering of the source language syntax (to align it closer to the target language word order) by Collins et al. (2010).

Furthermore, synchronous parsing (Henderson et al. 2008) will be used in the COMTIS project to improve current SMT models. Synchronous parsing provides a framework for jointly modeling multiple structures. For SMT, these structures are typically the syntactic parses for two sentences in different languages. To apply the parses in SMT, the more straightforward is a multi-version system in combination with a phrase-based SMT model, either interpolating between the two models or identifying when a literal translation is appropriate. The second approach provides more scope for advancing the state-of-the-art, and will be a synchronous model where two separate semantic structures are generated for the two sentences. Synchronous parsing can additionally provide insight to new features of use for Topic 3, which can, in return, provide new labels to the structures used by the synchronous parser.

2.5 Topic 5: Evaluation of improvement in MT coherence

Finally, since inter-sentential dependencies have scarcely been addressed, few metrics or evaluation methods are able to measure improvements on the intersentential and discourse levels. Topic 5 targets the creation of metrics and test-suites that are necessary to assess the correct translation of text cohesion. These will be used to evaluate the actual output of the SMT systems built in the project, and also to compare them with existing systems.

2.6 Overall approach

All the five topics described closely interact throughout the project. Figure 1 provides an example for Topics 1 to 4.

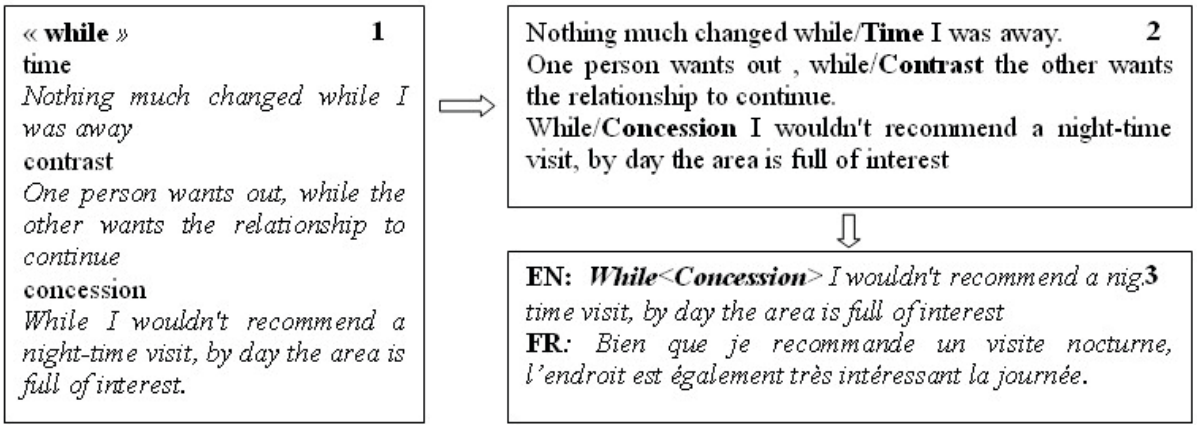

Figure 1: Steps in inter-sentential disambiguation for MT in COMTIS

In a first step (exemplified in box 1), diverging linguistic phenomena are identified, based on theoretical research, linguistic description (here, examples are extracted from the Oxford Online dictionary[1]) and empirical research on monolingual and bilingual corpora. In the example in Figure 1, three possible senses are thus identified for the connective "while". In a second step (shown in box 2), large number of occurrences are labelled with the different possible "meanings" identified in step 1. The annotation is first performed manually (see section 3 for different strategies adopted for this task) and then, the annotated corpus is used to train a classifier for further automatic annotation. Box 3 shows an example of a bi-sentence, where the English connective is automatically labelled (by the classifier). Annotated bi-texts are then used to train an SMT system.

As previously mentioned, topics 4 and 5 are in early stages so far. In the following, we present a case study based on three connectives (one for French and two for English), in the perspective of the three first topics of the project (linguistic analysis, corpus annotation and automatic annotation).

3. Connectives in French and English: a case study for the COMTIS Project

The COMTIS project started in March 2010. The first three topics have been under work, and the first results are briefly summarized here. In the following, we present the results of first experiments based on three connectives, two in English (*since* and *while*) and one for French (*alors que*).

3.1 Issues

Connectives are specific discourse markers that link two sentences together. One connective can have different senses, but it is rare that different languages share the same "polysemy" phenomena. For example, in the PDTB, the English connective *while* is annotated with more than twenty senses, because the annotators were allowed to choose senses from a three-level sense hierarchy using

(1) http://www.oxforddictionaries.com/

any possible combinations out of them. Thus, annotations for "while" result in a very broad sense distribution[2]. Out of them, four main senses can be distinguished (as in a preliminary experiment of the Penn Discourse Treebank annotation (Miltsakaki et. al. 2005)), which are: "opposition", "comparison", "concession" and "temporal", each of them having different translations into French.

Similarly, *since* conveys a temporal and a causative meaning (and in some contexts, both meanings can be present at the same time)[3], and should be translated differently into French.

Table 1 and Table 2 provide bi-sentences from the En-Fr Europarl Corpus (Koehn 2005) with two different meanings of *while* and *since*, that lead (correctly) to two different French translations (in bold in the text).

Table1: En-Fr while-sentences from the Europarl

While we have a duty to tackle this problem within EU waters, ultimately this is a problem which requires international action.	**Bien que** nous ayons le devoir de traiter ce problème au niveau des eaux de l'UE, il s'agit en dernier ressort d'un problème qui exige des actions au niveau international.
No wonder Richard Holbrooke recently boasted that Europe slept **while** President Clinton resolved a particular European crisis.	Il n'y a dès lors rien d'étonnant à ce que M. Richard Holbrooke nous ait récemment nargué en disant que l'Europe dormait **pendant que** le président Clinton résolvait une crise européenne particulière.

Table2: En-Fr since-sentences from the Europarl

In East Timor an estimated one-third of the population has died **since** the Indonesian invasion of 1975.	Au Timor oriental, environ un tiers de la population est décédée **depuis** l'invasion indonésienne de 1975.
Can I also touch on the question of taxation **since** Mr Görlach mentioned it in his speech.	Permettez-moi aussi de dire quelques mots en matière de fiscalité **puisque** M. Görlach en a parlé dans son intervention.

In French, the connective *alors que* conveys two meanings ("contrast" and "background"), according to the LexConn database (Roze *et al.* 2010), and similarly, it gives rise to different translations depending on the meaning, as shown in the two bi-sentences of Table 3, extracted from the Fr-En Europarl Corpus (Koehn 2005).

Table3: Fr-En alors que-sentences from the Europarl

Alors que l'Union soviétique a disparu et que le danger, plus diffus mais redoutable, vient aujourd'hui du Sud, les rivalités qui opposent les principales puissances occidentales se sont accentuées.	**Now that** the Soviet Union has disappeared and danger, less clear-cut but just as formidable, today threatens from the south, the rivalries between the major Western powers have become more pronounced.
La nouvelle négociation n'est pas enclenchée, **alors que** nous sommes à un mois de l'échéance	New negotiations have not been initiated, **even though** the deadline is only one month away

(2) Out of 781 occurrences of ***while***, the following senses (numbers of occ. in parenthesis) are observed in the PDTB data: COMPARISON (18), COMPARISON/Synchrony (4), Concession (1), Conjunction (39), Conjunction/Contrast (1), Conjunction/juxtaposition (5), Conjunction/Synchrony (21), Conjunction/TEMPORAL (1), contra-expectation (3), Contrast (120), Contrast/Synchrony (22), expectation (79), expectation/Synchrony (3), juxtaposition (182), juxtaposition/List (9), juxtaposition/Synchrony (26), List (3), List/opposition (1), opposition (78), opposition/Synchrony (11), Synchrony (154).

(3) The PDTB describes these three meanings as "reason", "succession" and "reason/succession".

The objective of the COMTIS project is to be able to automatically identify these different meanings and to label them, so that the SMT system can learn from these labels and disambiguate the connectives for translation. But to meet this objective, a first step for collecting and annotating data is required, through various annotation steps.

3.2 Manual Annotations

Following the methodology set up for the COMTIS project, the first step is to look at occurrences of the considered phenomena in corpora, and to try to perform human annotation that would disambiguate the meanings of such ambiguous items. In this vein, two different approaches have been adopted. First, a rather classic approach of "semantic" annotations has been adopted, making use of human annotators that were asked to assign a specific label to the different meanings of the connectives under consideration. As shown in the result below, this approach has shown some limitations in our case. A second approach, called translation spotting, was also tested, and proved to provide more fine-grained information about the actual meaning and use of connectives. Subsections below describe the two approaches.

3.2.1 The classic approach: semantic annotation of connectives

Large sets of monolingual sentences were given to two independent annotators that were asked to assign sense labels to a connective found in each sentence they saw. For example, for the French *alors que*, annotators could choose between "background" and "contrast", which were briefly defined at the beginning of the task (two other labels were also provided, one if the annotator could not decide, and one if *alors que* was not a connective, but a homographic character string).

The results show an inter-annotator agreement score (Cohen's Kappa) of 0.428, which is rather low, and shows that the task was not as simple as expected.

For *while*, two online surveys were carried out by five project members to annotate 30 sentences containing this connective, with the four main senses ('opposition', 'concession', 'comparison', 'temporal'). In experiments described by Miltsakaki et. al. (2005), the annotators agreed on 67 sentences out of 80. The most difficult distinction was the one between "opposition" and "concession". In our experiments, for the first 10 sentences, the inter-annotator agreement reached a kappa value of 0.60 (with 4 annotators), and for the other 20 sentences a kappa value of 0.56 (with 5 annotators). The highest disagreement was, in both queries, the distinction between "opposition" and "comparison".

Such annotation tasks thus appear to be more complex than expected. Another approach for annotating connectives has been adopted, and is now described.

3.2.2 The "translation spotting" approach: the case of "while"

Translation spotting consists of the manual annotation of the correspondences that have been used to translate a specific item. It is performed by human annotators on bilingual sentences pairs (though some attempts have been made to perform this task automatically (Huet et al. 2009) but proved to be particularly unreliable for function words such as connectives).

For the translation spotting task of *while* we used 508 bi-sentences extracted from the Europarl Corpus (Koehn 2005) for the English-French pairs, and we carefully extracted sentences that were originally stated in English (indeed, it has been shown in previous research that the use of connectives varies greatly in original or in translated text).

Two human annotators performed the tasks. They were asked to write down the connective that was used to translate the connective while. If it was not translated by a French connective, they were allowed to assign different tags (for the use of a present participle (G-tag), a paraphrase (P-tag), or no translation at all (Z-tag)).

Although the task might seem trivial, the two annotators provide a different translation spotting for 157 sentences out of the 508. Most of the mistakes are due to oversights, or to the rather unclear definition of connectives (when annotators confuse connectives like "s'il est vrai que" with paraphrases).

Table 4[4] provides details about the main different correspondences used to translate "while" in French, once the two "translation spotting" results have been manually merged and corrected by a third person.

Table 4: Translation spotting results for *while*

Translation of *while*	Nbr.	%
alors que	91	18.24 %
G	85	17.03 %
P	72	14.43 %
si	54	10.82 %
Z	41	8.22 %
tandis que	39	7.82 %
même si	33	6.61 %
bien que	26	5.21 %
s'il est vrai que	14	2.81 %
tant que	10	2.00 %
pendant	5	1.00 %
puisque	5	1.00 %
lorsque	4	0.80 %
mais	4	0.80 %
...	...	...
Total	**499**	100 %

As we can see, a wide range of French connectives is used to translate *while*. To deduce English meanings out of the translation, a supplementary step of "clustering" is needed, by analyzing and testing French connectives.

3.3 Clustering meaning out of translation

To distinguish the different meanings of *while* from the translation, we have to decide if the connectives used in French are "distinguishable" or are just "interchangeable" (and so conveying the same meaning). For the 6 most frequent French connectives used to translate *while*, (*alors que, si, tandis que, même si, bien que, s'il est vrai que*), we run a substitutability test, i.e. we test for a set of sentences which connective can be replaced by the other. (The same test was

(4) Among the 508 "while", 499 were actual connectives, the other being the noun "while", like in "for a while", or "a while ago", and have been excluded from the count. Table 4 does not provide figures for translation equivalents that appears less than 4 times

also performed for the temporal French connectives: *pendant, pendant que, tant que, puisque, lorsque, quand*).

These tests of substitutability allow us to cluster French connectives that convey the same meaning, and consequently narrow the different possible meanings of English *while*.

Table 5 summarizes the different meanings that have been found by clustering the French connectives that were seen to translate *while*. Only French connectives that appear at least twice have been kept.

Table5: clustered meanings for while

Translation of *while*	Token	%	Meaning
si	54	10.82 %	concession
même si	33	6.61 %	concession
bien que	26	5.21 %	concession
s'il est vrai que	14	2.81 %	concession
malgré	3	0.60 %	concession
quoique	3	0.60 %	concession
tandis que	39	7.82 %	contrast
mais	4	0.80 %	contrast
alors que	91	18.24 %	temporal/contrast
tant que	10	2.00 %	temporal/cause
puisque	5	1.00 %	temporal/cause
pendant	5	1.00 %	temporal/duration
lorsque	4	0.80 %	temporel/punctual

Table 5 shows that *while* displays more specific meanings than previously highlighted, particularly for the temporal one that seems to be more precise and detailed than expected.

Thanks to this clustering method of possible translation correspondence, we can now proceed to annotate the different meanings of *while* in the English sentences. This large amount of data can be further used for automatic annotation, as presented in the next section[5].

3.4 Automatic annotation

At the current state of the COMTIS project we could make use of basic surface features to disambiguate discourse connectives (see Section 3.3.2). In Section 3.3.3 we report the result for an experiment in automatically disambiguating the senses of *while* and *since*. This first classification was performed on the gold standard annotations of the PDTB, as the annotation and translation spotting tasks described above were still ongoing. It was also important to first focus on the PDTB data in order to re-employ some of the features used by current research and the state-of-the-art for discourse connective classification. This also helps to make our results comparable. In addition, the PDTB data is linked to the Penn Treebank gold parses, which guarantees the feature's robustness in these early stages of disambiguation. For a description of the PDTB, see the following section 3.3.1.

(5) Although the automatic annotation experiment presented here does not make use of the same material, because this experiment in particular predates the translation spotting result.

3.4.1 The Penn Discourse Treebank

The PDTB is one of the very few available discourse annotated resources. There are 100 types of explicit connectives annotated with their senses.[6] The sense hierarchy used in the PDTB consists of three levels, from four top level senses through 16 subsenses on the second level and 23 further subsenses on the third level (see the PDTB annotation manual for a full description of the hierarchy[7]).

The PDTB further sees the connective as a discourse-level predicate that has two arguments. Argument2 is the one containing the explicit connective. As an example, the sentence from Table 1 above could be represented as *while(ultimately this is...[argument1], we have a... [argument2])*.

3.4.2 Features

So far, we implemented the following features: (1) the connective word form, (2) its POS tag, (3) argument1's first word, (4) argument1's last word, (5) argument2's first word (6) argument2's last word, (7) argument2's first word's POS tag, (8) type of argument2's first word, (9) parent syntactical categories of the connective, (10) punctuation pattern.

The cased connective word forms in the PDTB (feature 1) were left as is, therefore also indicating whether the connective is located at the beginning or in the middle of a sentence. The variations from the PDTB (e.g. *since – ever since* etc.) were also included, supplemented by their POS tags (feature 2). As shown by Lin et al. (2010) and duVerle and Prendinger (2009), the context of a connective is very important. The arguments may include other (reinforcing or opposite) connectives, numbers (for numerical comparison) and antonyms (to express contrastive relations (*rise* vs *fall*)). We extracted the words at the beginning and at the end of argument1 (features 3, 4) and argument2 (features 5, 6) which likely are other connectives, gerunds, adverbs or determiners (further generalized by features 7 and 8). The paths of syntactical ancestors (feature 9) in which the connective appears are quite numerous and therefore truncated to a maximum of four ancestors (e.g. |SBAR||VP||S|, |ADVP||ADJP||VP|, etc). Punctuation patterns (feature 10) are of the form C,A – A,CA etc. where C is the explicit connective and A a placeholder for all the other words. Punctuation is important for locating connectives as many of them are subordinating and coordinating conjunctions, separated by commas (Haddow, 2005).

3.4.3 Disambiguation results for *since* and *while*

For the connective *since* there is a total of 150 token occurrences in the PDTB training set.[8] If the corresponding senses are reduced to 'temporal' and 'cause' only, there are 83 occurrences of *since* with a causal meaning and 67 with a temporal meaning.

For *while*, the 21 senses in the PDTB can be reduced to the following four: contrast, concession, temporal and expansion. The 631 occurrences of *while* in the training set have the following distribution of senses: 342 contrast, 159 temporal, 77 concession, 53 expansion.

For disambiguation of connectives, we report here results based on 10-fold cross validation on the training set for *since* and the one for *while*. As classifier we used the implementation of the RandomForest decision tree algorithm in the WEKA machine learning toolkit (Hall et al., 2009)[9].

(6) There are also implicit relations, for which the annotators had to guess a connective most probably fitting in between two text spans related without a lexically explicit cue word.
(7) http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf
(8) The PDTB training set consists of sections 02-22
(9) http://www.cs.waikato.ac.nz/ml/weka/

Disambiguating the two senses 'temporal' and 'cause' for *since* with this classifier leads to an accuracy of 75.3% of correctly classified instances, which is significantly above the baseline of 55.3% (prediction of the majority class (the sense "cause")).

For *while*, disambiguating its 4 senses leads to an accuracy of 59.6%, which is also significantly above the baseline of 54.1% (prediction of the majority class "contrast").

Possible problems and points to consider in further research are additional and better features, for example, including polarity features and semantic relations (antonyms). The experiments will then be extended to the other project languages as well as to other sets of explicit connectives.

4. Conclusion and future work

In this paper, we presented the outline of the COMTIS project that aims at improving the textual coherence in machine translation. We described the different topics covered by the project that try to bring together methods and techniques from theoretical linguistics, corpus-based analysis and natural language processing.

Until now, most of the research work focused on the annotation of linguistic phenomena in order to disambiguate them. Annotating connectives proved to be a difficult and complex task, and different methodologies have been put in place to address it. Namely, the "translation spotting" method seems to shed new light on distinctions that are necessarily made in monolingual studies or annotation frameworks.

As mentioned, the project is at its early stage, and further works is planned such as large-scale annotation (specifically translation spotting) of connectives, precise definition of features to annotate verbal tenses discrepancies, and the resolution of anaphora.

Bibliographie

Carpuat, M., Wu, D., (2005). "Word sense disambiguation vs. statistical machine translation". In: ACL 2005 (43rd Annual Meeting of the Association for Computational Linguistics). Ann Arbor, MI, USA, pp. 387-394.

Carpuat, M., Wu, D., (2007). "Improving statistical machine translation using word sense disambiguation". In: EMNLP-CoNLL 2007 (Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning). Prague, Czech Republic, pp. 61-72.

Collins, M., Koehn, P., Kucerova, I., 2005. Clause Restructuring for Statistical Machine Translation. In: Proceedings of the 43rd Annual Meeting of the ACL.

Dorr, B. J., Gaasterland, T., 1995. Selecting tense, aspect, and connecting words in language generation. In: IJCAI 1995 (14th International Joint Conference on Artifcial Intelligence). vol. 2. Montreal, pp. 1299-1307.

Dorr, B. J., (1992). "A parameterized approach to integrating aspect with lexical-semantics for machine translation". In: ACL 1992 (30th Annual Meeting of the Association for Computational Linguistics). Newark, DE, USA, pp. 257{264.

Dorr, B. J., (1993). Machine Translation: A View from the Lexicon. The MIT Press, Cambridge, MA, USA.

Dorr, B. J., Gaasterland, T., (1995) "Selecting tense, aspect, and connecting words in language generation". In: IJCAI 1995 (14th International Joint Conference on Artificial Intelligence). vol. 2. Montréal, pp. 1299-1307.

duVerle, D., Prendinger, H., (2009) A Novel Discourse Parser Based on Support Vector Machine Classification. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP.

Hall M., Frank E., Holmes G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1).

Grote, B., Stede, M., (1998) "Discourse marker choice in sentence planning". In: INLG 1998 (9th International Workshop on Natural Language Generation). Niagara-on-the-Lake, ON, pp. 128-137.

Haddow, B., 2005. Acquiring a Disambiguation Model For Discourse Connectives. Master Thesis. University of Edinburgh, School of Informatics.

Henderson, J., Merlo, P., Musillo, G., Titov, I., 2008. "A latent variable model of synchronous parsing for syntactic and semantic dependencies". In: Proceedings of CONLL 2008. Manchester, UK, pp. 178-182.

Hovy, E. H., (1988) "Generating Natural Language Under Pragmatic Constraints". Lawrence Erlbaum Associates, Hillsdale, NJ.

Huet S., Bourdaillet J. and Langlais L. « Intégration de l'alignement de mots dans le concordancier bilingue TransSearch ». Proceedings of TALN'09. Senlis, France, June 2009

Hutchinson, B., (2004) "Acquiring the meaning of discourse markers". In: Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics). Barcelona, Spain, pp. 685-692.

Koehn P. (2005) "Europarl: A Parallel Corpus for Statistical Machine Translation", MT Summit 2005

Lapata, M., Lascarides, A., (2004) "Inferring sentence-internal temporal relations". In: HLT-NAACL 2004 (Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics). Boston, MA, pp. 153-160.

Le Nagard, R., Koehn, P., 2010. Aiding Pronoun Translation with Co-Reference Resolution. In: Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR.

Lin, Z., Ng, H. T., Kan, M., 2010. A PDTB-styled End-to-End Discourse Parser. Technical Report TRB8/10. School of Computing, National University of Singapore.

Litman, D. J., (1996) "Cue phrase classification using machine learning" Journal of Artificial Intelligence, Research 5, 53-94.

Mann, W. C., Thompson, S., (1988) "Rhetorical structure theory: toward a functional theory of text organization". Text 8 (3), 243-281.

Marcu, D., (2000) "The Theory and Practice of Discourse Parsing and Summarization". MIT Press, Cambridge, MA.

Miltsakaki E., Dinesh N., Prasad R., Joshi A., Webber B. (2005). Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT).

Mitkov, R., (2002) "Anaphora Resolution. Studies in Language and Linguistics". Longman, London, UK

Morris, J., Hirst, G. (1991) "Lexical cohesion computed by thesaural relations as an indicator of the structure of text". Computational Linguistics 17 (1), 21-48.

Murata, M., Nagao, M., (1993) "Determination of referential property and number of nouns in Japanese sentences for machine translation into English". In: TMI 1993 (5th Conference on Theoretical and Methodological Issues in Machine Translation). Kyoto, Japan, pp. 218-225.

Murata, M., Uchimoto, K., Ma, Q., Isahara, H., (2001) "A machine-learning approach to estimating the referential properties of japanese noun phrases". In: CICLing 2001 (2nd International Conference on Computational Linguistics and Intelligent Text Processing). Mexico-City, Mexico, pp. 142-153.

Nakaiwa, H., Ikehara, S., (1992) "Zero pronoun resolution in a machine translation system by using Japanese to English verbal semantic attributes". In: ANLP 1992 (Conference on Applied Natural Language Processing). pp. 201-208.

Nakaiwa, H., Ikehara, S., (1995) "Intrasentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints". In: TMI 1995 (6th International Conference on Theoretical and Methodological Issues in Machine Translation). Leuven, Belgium, pp. 96-105.

Nakaiwa, H., Shirai, S., (1996) "Anaphora resolution of Japanese zero pronouns with deictic reference". In: COLING-96 (16th International Conference on Computational Linguistics). Copenhagen, pp. 812-817.

Peral, J., Palomar, M., Ferrandez, A., (1999) "Coreference-oriented interlingual slot structure and machine translation". In: ACL 1999 Workshop on Coreference and Its Applications. College Park, MD, USA, pp. 69-76.

Pitler, E., Nenkova, A., 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B. (2008) The Penn Discourse Treebank 2.0. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).

Roze C., Danlos L. & Muller Ph. (2010) "LEXCONN: a French Lexicon of Discourse Connectives" Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010), Moissac.

Reichenbach H. (1947), 'The tenses of Verbs'. Section 51 of *Elements of Symbolic Logic,* The McMillan Company: New York, pp. 287-298

Sperber D., Wilson D. (1986) *Relevance: Communication and Cognition.* Basil Blackwell.

Vendler Z. (1957), 'Verbs and times'. In *The Philosophical Review*, vol. 66, No. 2, pp. 143-160

Wilde, O. (1889) "The portrait of Mr. W.H.", Editions Gallimard, 2000. The *JRC-Acquis Multilingual Parallel Corpus* http://langtech.jrc.it/JRC-Acquis.html