A Corpus-based Contrastive Analysis for Defining Minimal Semantics of Inter-sentential Dependencies for Machine Translation

Thomas Meyer, Andrei Popescu-Belis, Jeevanthi Liyanapathirana
Idiap Research Institute, Martigny, Switzerland
Bruno Cartoni
University of Geneva, Switzerland

Abstract

Inter-sentential dependencies such as discourse connectives or pronouns have an impact on the translation of these items. These dependencies have classically been analyzed within complex theoretical frameworks, often monolingual ones, and the resulting fine-grained descriptions, although relevant to translation, are likely beyond reach of statistical machine translation systems. Instead, we propose an approach to search for a minimal, feature-based characterization of translation divergencies due to inter-sentential dependencies, in the case of discourse connectives and pronouns, based on contrastive analyses performed on the Europarl corpus. In addition, we show how to automatically assign labels to connectives and pronouns, and how to use them for statistical machine translation.

1. The Need for Inter-sentential Information in Machine Translation

Long-range dependencies are a well known challenge for machine translation (MT) systems, especially for statistical ones. The correct translation of lexical items such as pronouns often depends on the correct identification of their antecedent. Similarly, the correct translation of multi-functional discourse connectives depends on the correct identification of the rhetorical relation which they convey between two clauses. However, especially when translating between closely related languages, the full disambiguation of such lexical items is sometimes unnecessary for a correct translation. The question that arises is thus how to find the most suitable level of representation for such dependencies, as a trade-off between linguistic accuracy and computational tractability, with the direct aim of improving MT output.

This paper presents a method for finding the minimal semantic/discourse information that must be assigned to two types of lexical items, namely connectives and pronouns, in order to avoid translation mistakes by statistical MT systems. The method starts from contrastive analyses of a frequently used parallel corpus, Europarl (Koehn, 2005), in order to define and annotate the minimal semantic/discourse information necessary for MT. The paper first describes our analyses and manual annotation methods for disambiguating connectives (Section 2.1) and pronouns (Section 2.2), in the context of English/French MT. Section 3 outlines methods for automatically performing these disambiguation tasks, while Section 4 explains how the automatically labeled linguistic items can be integrated into a statistical MT system. Section 5 concludes the paper and outlines future work.

2. Contrastive Analysis of Two Types of Inter-sentential Dependencies

2.1 Discourse Connectives

Discourse connectives are generally considered as indicators of discourse structure, relating two sentences or propositions and making explicit the rhetorical relation between them. Explicit discourse connectives such as *because, but, however, since, while*, etc., are frequent

lexical items and are used to mark rhetorical relations such as *Cause* or *Contrast* between units of discourse. Several theoretical frameworks have been proposed for connectives (mainly starting from English ones), such as the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), or the Segmented Discourse Representation Theory (SDRT) (Asher, 1993). In such theories, more than one hundred possible rhetorical relations have been identified, and complex semantic and logical representations have been used to characterize discourse structure. In a more empirically oriented effort, the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) contains manual annotations of discourse connectives with a large set of labels: for example, the connective *while* was annotated with 17 possible senses beyond its for main meanings, which are *Comparison*, *Contrast*, *Concession* and *Opposition* (Miltsakaki et al., 2005).

While a fine-grained characterization provides the necessary theoretical level of linguistic description of discourse structure, it may prove to be intractable to fully automatic processing. Nevertheless, the disambiguation of at least the main senses of discourse connectives is generally required for their translation[1], to avoid the rendering of a wrong sense in translation. For instance, in the following example, the French connective *alors que* in its *contrastive* usage is wrongly translated to the English connective *so*, which signals a *causal* meaning instead[2].

FR: *Oui, bien entendu, sauf que le développement ne se négocie pas,* ***alors que*** *le commerce, lui, se négocie.*

EN: **Yes, of course, but development cannot be negotiated,* ***so*** *that trade can.*

To disambiguate connectives for MT, parallel corpora with sense-labeled connectives are required for training and test. As the PDTB data is in English only, we performed manual annotation on the Europarl corpus. The annotation method, called translation spotting, requires annotators to consider bilingual sentence pairs, and annotate each connective in the source language with its translation in the target language (Meyer et al., 2011). A contrastive analysis showed that these translations can be: a target language connective (in principle signaling the same sense(s) as the source language one), reformulations with different syntactical constructs, or no connective at all. The indications gained with this method are then used in a second step to manually derive and cluster the minimal semantic and theory-independent labels needed to generate correct translations of a connective.

We exemplify this procedure here for the English connective *while*. From the Europarl corpus for English-French, we extracted 499 sentences containing the connective *while*. In 198 cases (43%) the annotators spotted 'no translation' or reformulations of the connective[3]. In the remaining 301 sentences (57%), the annotators identified the corresponding French connectives. As a second step, the French connectives (signaling the same rhetorical relation(s) as *while* itself) were manually clustered under the minimally necessary sense labels to disambiguate the connective *while* in order to translate it correctly from EN to FR. The most frequent French connective clusters and the derived sense labels are the following:

alors que (18%)	Contrast/Temporal
si / même si / bien que / s'il est vrai que (25%)	Concession
tandis que / mais (9 %)	Contrast
tant que (2%)	Temporal/Causal
pendant (1%)	Temporal/Duration
puisque (1%)	Temporal/Causal
lorsque (0.8%)	Temporal/Punctual

[1] The only exception is the case when the ambiguity of a connective is conserved in translation.
[2] Source sentence from Europarl, translated by Moses (Koehn et al., 2007) trained on Europarl.
[3] These are valid translation problems and will be reconsidered for clustering in future work.

Compared to the PDTB sense hierarchy for example, the clustered senses for *while* are as detailed as the PDTB ones on hierarchy level 2, but less detailed than the deepest PDTB level 3. For the temporal meaning of *while,* however, even more differentiation than PDTB level 3 is needed in order to be able to generate the correct translations.

2.2 Pronouns

The resolution of pronouns can be seen as a similar issue to that of resolving connectives in terms of finding a minimal set of features to disambiguate a pronoun for translation. In many cases, depending on the language pair, pronouns can be translated unequivocally, such as the English pronoun *he* generally rendered by *il* in French. However, the French pronouns *il* and *elle* may both be translated into *it* in English if their antecedent, i.e. the noun they refer to, is not human. However, if the antecedent is human, they are in general translated respectively as *he* and *she*. Vice versa, the translation of the English pronoun *it* into French requires knowledge about the gender of its antecedent in the target text. Therefore, whereas the disambiguation of connectives can be done on the source text only, prior to MT, the translation of pronouns requires information about the translation of neighboring fragments.

A close comparison of the English and French pronoun systems shows that the complete list of features characterizing pronoun choice is in theory very large. However, we only aim here to find the minimal set of features which will allow a statistical MT system to avoid generating mistaken pronouns, taking also into consideration the pronoun generated by the system without these features. For instance, in the following example from Europarl, the pronoun generated by Moses is correct in every respect except the gender; therefore, knowledge about the required gender would help correcting *il* into *elle*.

EN: *The **European Commission** must make good these omissions as soon as possible. **It** must also cooperate with the Member States...*

FR: **La **Commission européenne** doit réparer ces omissions dès que possible. **Il** doit également coopérer avec les États membres ...*

3. Automated Disambiguation for Machine Translation

To improve the output of MT, we propose automatic methods that attempt to disambiguate, or at least set additional constraints, on the translation of connectives and pronouns. These methods can either be used as direct input to MT, or to prepare training data for it. For instance, using surface features such as part-of-speech tags or syntactical and dependency parses, we have built classifiers (Meyer et al., 2011) for the senses of the English connectives *since* (Temporal, Causal, or Temporal/Causal) and *while* (Temporal/Causal, Temporal/ Punctual, Temporal/ Durative, Contrast/Temporal, Contrast, or Concession), as well as for the French connective *alors que* (Temporal, Contrast, Temporal/Contrast).

	since	*while*	*alors que*
Baseline (most frequent sense)	51.6%	44.8%	46.9%
SVM classifier	85.7%	60.9%	54.2%

Table 1: Accuracies of sense disambiguation for the connectives *since* (700 sentences), *while* (300) and *alors que* (400). For comparison, the baseline is the majority class in each training set, i.e. respectively *Cause, Concession,* and *Contrast*.

Classifiers were also built for pronoun disambiguation, considering in addition to features from the source text also features from a candidate translation, such as information about the preceding noun phrases, the candidate Moses translation of the pronoun computed from the GIZA++ word alignment, and various ways to determine gender constraints – for the translation of English *it* into French – from the gender of the preceding nouns (e.g., majority,

most recent, etc.). Although this method bears similarities with that of LeNagard and Koehn (2010), we do not attempt to identify explicitly the antecedent, in the target language, of the pronoun under consideration, but train classifiers to use the optimal combination of features to infer the correct gender. Of course, this approach cannot pretend to be fully accurate, but compares favorably to state-of-the-art accuracy of automatic pronoun resolution.

The accuracy of the classifier, a decision tree trained using the C4.5 algorithm, is 61% using ten-fold cross-validation on a set of 393 sentences from Europarl annotated with the correct pronoun. The task was to correct the Moses candidate translation of English *it* into French (*il, elle, le, la, l', lui, celui-ci, celle-là, ce, c'*) using automatic alignment and automatically extracted surface features. If the alignment is manually corrected, then the accuracy reaches 64%. This small increase shows that alignment is not the main issue, also because it cannot deal with cases when the MT system omitted the pronoun in translation. However, when the gender prediction is manually corrected, the accuracy reaches 88%, which shows that, as expected, gender is the main feature required for correct translation of *it* into French.

4. Integration into Statistical MT

We experimented on three ways to propagate the above-mentioned discourse information annotated to connectives into the MT processing chain. The integration of annotated pronouns proceeds differently, as a way to post-edit candidate pronouns generated by MT.

The first method to integrate the minimal sets of labels for discourse connectives is to tag their occurrences directly in the phrase table of an already trained statistical MT system. During the training stage, a phrase table is generated with all phrase pairs found by the word alignment, with their lexical probability and frequency scores. We tagged three senses of the connective *while*, namely *Temporal* (1), *Contrast* (2) and *Concession* (3) in the phrase table of a trained Moses MT system for EN-FR. The most frequent French translations were: (1) *pendant que*, *(tout) en + V-ant*, (2) *alors que, tandis que*, (3) *bien que*. Each phrase containing *while* was automatically checked if it is followed by a corresponding translation. If found, the word form *while* was annotated with *while-1*, *while-2* or *while-3*, and, in addition, the lexical probability score was set to one (all other occurrences were left untagged). Translations tests with a set of 20 sentences already led to noticeably better translations (i.e. automatically generated translations closer to the reference translations, especially in terms of the connective) which were also confirmed by a rise in the BLEU score of 0.8 absolute.

A second method that we explored is the opposite of forcing the system to use the tagged connectives. They are instead automatically tagged in a large corpus which is used for SMT training, where all connectives followed by their tags and their corresponding translation in the parallel corpus can be learned by the system. Every occurrence has thereby to be tagged by the disambiguation tool using the classifier model. A third and similar approach to this method is to directly use the manually annotated discourse connectives after the sense clustering. This has the advantage that the hand-annotated resources are correct (gold standard) as opposed to the automated tagging, which is well below 100% accuracy and may therefore propagate a certain error rate in the whole translation process. We built and trained SMT systems able to handle the same manually or automatically tagged data. As a basis for comparison, two other systems were trained on the same two corpora, by discarding all labels (resulting in 4 SMT systems). When comparing the manually tagged system to its untagged counterpart, the tagged system got closer to the reference translations of a test set of 35 sentences in 21 cases versus 14 cases only for the untagged system (the counts were done based on manual checking of the connective translation and the surrounding words and

syntax). Even the automatically tagged system, tested on 62 sentences, performed noticeably better in 14 cases compared to its untagged counterpart.

For pronouns, we evaluated the effect on translation of replacing every candidate translation of the English *it*, in the MT output to French, by the translation proposed by our classifier, as a form of post-editing. By definition, this method is only applicable to sentences where a pronoun was indeed generated by MT (about 95% of the sentences). We performed five different runs, training on 353 sentences and testing on 40. In the fully automatic setup, this resulted, on average, in improving pronoun choice from incorrect to correct in 10.8 sentences (27%), but also in turning 6.6 (16%) correct pronouns into incorrect ones. The global result is thus an improvement of about 10% of the overall pronoun accuracy. In these experiments, our classifier did not change the pronoun proposed by MT in 22.6 sentences (56%), of which 27% were correct and 29% were incorrect.

5. Conclusion and Future Work

Integrating discourse information into statistical MT systems remains a challenging task, but one which has the potential to improve over the current sentence-by-sentence MT paradigm. The contrastive corpus analyses and the translation-oriented, multilingual annotation methods have shown to positively affect the output of current statistical MT systems. We will further investigate the automated disambiguation methods for pronouns and connectives as well as for verbal tenses. The performance and error rate of the disambiguation tools is crucial in order to generate annotated resources which are as error-free as possible in order to not negatively influence the SMT training and testing on these resources.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publisher, Dordrecht, NL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007 (45th Annual Meeting of the ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pp. 79–86, Phuket, Thailand.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: towards a functional theory of text organization. *Text*, 8(3):243–281.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. *Proceedings of SIGDIAL 2011 (12th annual SIGdial Meeting on Discourse and Dialogue)*, pp. 194–203, Portland, OR.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the TLT 2005 (4th Workshop on Treebanks and Linguistic Theories)*, Barcelona, Spain.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, pp. 258–267, Uppsala, Sweden.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008 (6th International Conference on Language Resources and Evaluation)*, pp. 2961–2968, Marrakech, Morocco.