discours
Revue de linguistique, psycholinguistique et informatique
http://discours.revues.org/

Discourse in Statistical Machine Translation

A Survey and a Case Study

Christian Hardmeier

Christian Hardmeier, « Discourse in Statistical Machine Translation », *Discours* [En ligne], 11 | 2012, mis en ligne le 23 décembre 2012.

URL : http://discours.revues.org/8726

Titre du numéro : *Varia*
Coordination : Olivier Ferret et Nicolas Hernandez

Discourse in Statistical Machine Translation

A Survey and a Case Study

Christian Hardmeier

Current approaches to statistical machine translation assume that sentences in a text are independent, ignoring the property of connectedness present in virtually all discourse. We provide an extensive overview of the literature about statistical machine translation that can be related to discourse phenomena and present a detailed investigation and discussion of existing research efforts on a particular discourse-related problem, the translation of anaphoric pronouns. Comparing different approaches to discourse in statistical machine translation allows us to identify fundamental problems and draw conclusions from an overarching perspective.

Keywords: statistical machine translation, discourse, pronominal anaphora, survey

1. Introduction

1 Machine translation (MT) researchers widely agree that translation is a complex task that cannot be solved by looking at words and their immediate local neighborhood only; however, much of the existing work on MT depends on strong assumptions of locality for practical reasons. The first models of the modern statistical machine translation (SMT) paradigm published in the early 1990s (Brown et al., 1990 and 1993) impose strong independence assumptions on the words in a sentence and only take into account a very limited context consisting of the one or two immediately preceding words in the target language for each word that the system outputs. Phrase-based and syntax-based SMT, the two currently dominant paradigms, relax these independence assumptions by considering a greater number of local dependencies, also in the source language. In syntax-based MT, some long-range dependencies inside the sentence can be accommodated. However, even advanced MT systems still assume that texts can be translated sentence by sentence and that the sentences in a text are strictly independent of one another.

2 From a linguistic point of view, this is unsatisfactory. A text to be translated is "more than a random set of utterances". Like all forms of written or spoken discourse, "it shows connectedness" (Sanders & Pander Maat, 2006: 591). Connectedness implies dependencies between sentences; if the dependencies are neglected in translation, there is a risk that the output text no longer has the property of connectedness which makes a sequence of sentences a text. In the field of translation studies, the importance of discourse was recognized long ago (Hatim & Mason, 1990). In SMT, it has only recently entered the focus of some research groups, and making use of discourse features to improve MT has turned out to be a daunting challenge. This article reviews existing responses to this challenge in a fairly extensive survey of the work published in the SMT literature that deals with cross-sentence context,

cohesion, discourse and other phenomena that can be related to discourse-level structures. We then zoom in to a specific case study and address the problem of translating pronominal anaphora, presenting and discussing previous work and identifying a number of problems that should be considered in future research.

2. Discourse in SMT: a survey

2.1. Discourse-related research in SMT

3 Current MT technology can be roughly categorized into two approaches, rule-based and corpus-based technology. Rule-based systems have been popular since the early days of MT research. They rely on hand-written translation rules, enabling them to capture complex relations between the input and the output language with high precision at the cost of considerable implementation and maintenance effort.

4 SMT, on which we focus in this article, is a corpus-based approach to MT: SMT systems derive statistical models from large collections of texts translated by human translators, for which both the input text and the translation are available (parallel corpora). The models estimated on a training corpus can then be used to generate translations for previously unseen documents that are not contained in the training corpus. SMT models have different components, the two most important model types being translation models (phrase tables, rule tables), which map linguistic entities from the source to the target language, and language models, which model target language fluency. Currently, there are two main approaches to SMT, both of which are capable of delivering state-of-the-art performance: in phrase-based SMT (Koehn, Och & Marcu, 2003), the translation units are contiguous sequences of a small number of words (not necessarily linguistic phrases); in hierarchical and syntax-based SMT (Chiang, 2007), translation is modeled with the help of a synchronous context-free grammar. The reader is referred to the recent textbook by Koehn (2010) for a more detailed introduction.

5 In the rule-based MT community, the use of discourse-level features was discussed already in the 1990s (e.g., Mitkov, 1999). In SMT, there has been a strong tendency to assume strict independence between the sentences in a document until recently. One of the earliest explicit references to discourse in the SMT literature dates from 2000. In a corpus study on discourse-level differences between English and Japanese, Marcu, Carlson and Watanabe (2000) announced that they were working on a discourse-enabled SMT system combining a discourse parser and a rule-based discourse transfer module with a SMT component. To our knowledge, this system has never been released or described in more detail. After that, we do not know of any explicit attempts to connect SMT with discourse until the work of Carpuat (2009) on the "one translation per discourse" hypothesis was published.

6 We believe that the reason for the SMT community's apparent lack of interest in discourse lies in the fact that the methods used in SMT have generally been

based on word-level pattern recognition and transformation and, unlike rule-based MT, have worked at a fairly low level of linguistic abstraction. Much existing work on discourse modeling is geared towards creating representations of texts in some abstract framework (Stede, 2011). While the degree of linguistic abstraction used by different rule-based systems varies (Isabelle & Foster, 2006; see, e.g., Forcada et al., 2011, for a recent example of a rule-based MT system with relatively shallow analysis), their architecture generally does include a sequence of analysis, transfer and generation steps that are explicitly modeled by human experts. The system response is tightly coupled to the rules, and targeted rules using discourse context can be added when it is considered useful.

7 The most successful SMT systems of the last two decades, by contrast, have rarely used any linguistic analysis more sophisticated than what is needed to split text into words (tokenisation/word segmentation). SMT embraces a low-level approach that considers word tokens only and eschews most linguistic abstractions. Empirically, this stance is justified by the fact that phrase-based SMT, whose core entities are word sequences unconstrained by any linguistic theory, and hierarchical SMT, which applies a purely formal and non-linguistic form of synchronous context-free grammars, often rival or beat the performance of approaches with stronger linguistic foundations. This is evidenced by the fact that many of the top-performing systems in recent MT shared tasks (e.g., Callison-Burch et al., 2012) use explicit linguistic models sparingly if at all.

8 The absence of strong linguistic models in SMT means that SMT cannot rely on linguistic intuitions to guide translation. Instead, the translation process is designed in a very generic way: in a phrase-based SMT system, translations of ambiguous words are selected by considering just the surrounding words, and the phrases in the output can a priori be permuted at will. Biases are then introduced on technical grounds to make model training and decoding (translating) practical: the order of the n-gram language model and the length of the phrases in the translation model are limited to sequences of just a few words, and so is the range of the permutations that are actually admitted for word reordering. In practice, this causes a very strong bias towards retaining all features of the input text that are not extremely local. This approach has turned out to be very hard to beat.

9 Even though explicitly discourse-related research topics became popular in the research community only very recently, this does not mean that there were no relevant publications at all in the years before. However, most of the earlier work that may be considered relevant to discourse treatment in SMT was written from different points of view such as domain adaptation or language modeling, and its connection to discourse may be less obvious. In the remainder of this section, we endeavor to give an overview of literature published by the SMT community that researchers working on discourse problems within this MT framework should be aware of. We do not attempt to describe advances in discourse modeling that are unrelated to SMT; other recent studies such as M. Stede's textbook on discourse processing

(Stede, 2011) or the survey article on discourse structure and language technology by Webber, Egg and Kordoni (2011) do this much better than we can. We have also chosen to restrict ourselves to research performed in the SMT framework without covering earlier publications about discourse in rule-based MT, some of which date back considerably earlier than the work discussed in our paper. We feel that the differences between the two approaches, especially with respect to the use of higher-level information, are so big that it is very difficult to transfer the insights and solutions described in the rule-based MT literature to SMT research, with the possible exception of the corpus studies occasionally included in the MT literature.

2.2. Lexical choice, consistency and context

2.2.1. Accessing context with word sense disambiguation

10 The first explicit attempts to use a larger context in SMT aimed at integrating word sense disambiguation (WSD) systems into SMT. WSD disambiguates the senses of polysemous words by considering the context in which they occur. Carpuat and Wu (2005) obtained negative results when they tried to combine a general Chinese WSD system with a word-based SMT system using IBM model 4. Vickrey et al. (2005) showed that a WSD component using word alignments from a bilingual corpus as sense annotations outperformed a monolingual language model in a simplified word prediction task simulating certain aspects of MT decoding. Later research suggests that WSD can have a positive effect when used to disambiguate complete phrases (as defined by the SMT system) if WSD scores are added as an additional context-dependent feature function to a phrase-based (Carpuat & Wu, 2007a and b) or hierarchical (Chan, Ng & Chiang, 2007) SMT decoder or by using local Support Vector Machine classifiers for phrase selection in a phrase-based SMT system (Giménez & Màrquez, 2007). Specia, Sankaran and das Graças Volpe Nunes (2008) showed that WSD-guided n-best-list reranking can improve the translations of a small number of highly ambiguous verbs in a treelet SMT system, but they only provided figures for a strongly biased test set in which every sentence contained at least one of the target verbs. When viewed in combination, the evidence suggests that just plugging together out-of-the-box components for different subtasks is insufficient, but that good results can be achieved when all the system components are integrated smoothly and modeled in terms of the same entities (SMT phrases, with aligned parallel text as sense annotations, in this case). Most of the papers cited above are not specific about the size of the context window taken into account by the WSD system, so it is unclear to what extent cross-sentence information was actually used for WSD, but the experimental setups would have permitted doing so.

2.2.2. Corpus studies on lexical cohesion and SMT

11 The integration of WSD into SMT can be seen not only as a means of ensuring more precise translations, but also as a factor potentially contributing towards improved lexical cohesion by favoring consistent vocabulary use throughout the document. Carpuat (2009) examined a specific kind of vocabulary consistency in translated texts

in the form of the "one translation per discourse" hypothesis, an assumption based on the well-known "one sense per discourse" hypothesis that is commonly used in language processing (Gale, Church & Yarowsky, 1992). By examining human reference translations for two English-French SMT test sets, she found indeed that 80% of the French words were linked to no more than one English translation and 98% to at most two translations, after lemmatizing both source and target. Looking at machine translations of the same test sets, she found that the regularity in word choice was even stricter in SMT as a result of the generally low lexical variability of SMT output.

12 These results suggest that there is not much to be gained by just enforcing consistent vocabulary choice in SMT, since the vocabulary is already fairly consistent. In principle, it may be possible to improve SMT by using whole-document context to select translations. However, a more recent study by Carpuat and Simard (2012) showed that this may be more difficult than it seems. In this study, the authors found consistency and translation quality to be essentially uncorrelated or even negatively correlated in SMT output. In particular, they showed that machine-translated output tended to be more consistent when produced by systems trained on smaller corpora, indicating that "consistency can signal a lack of coverage for new contexts" rather than being a sign of translation quality (Carpuat & Simard, 2012: 446). In a manual analysis of MT output post-edited by professional translators, they found that most of the inconsistencies observed were symptoms of more fundamental problems such as outright semantic translation errors or syntactic or stylistic problems, whereas the terminological inconsistencies typically found in imperfect human translations only accounted for about 13-16% of the inconsistent translations. These findings are encouraging in the sense that, in the best case, a model improving MT output consistency in the right way might help to fix some of the more fundamental errors as well, but the lack of positive correlation between measured consistency and translation quality shows that it is important to enforce not only consistent, but also correct translations, and that it may be necessary to make use of additional information for good results.

13 The "one translation per discourse" hypothesis was tested again by Ture, Oard and Resnik (2012), using a methodology based on forced decoding with a hierarchical SMT system and examining the translations selected by human translators at text positions where multiple options would have been available in the SMT rule table. They found that human translators indeed opt for consistent lexical choices in the majority of cases, but that some content words may be translated in more varied ways because of stylistic considerations. They proposed a set of cross-sentence feature functions rewarding translation rule reuse that achieved significant improvements in Arabic-English and Chinese-English translation tasks.

14 Yet another corpus study about lexical cohesion in MT output was published by Voigt and Jurafsky (2012). They compared referential chains in a literary text and a piece of news text in Chinese with their English translations generated by

Google Translate. They observed that, in the source language, both texts featured a similar number of entities, but that the referential chains in the literary text were denser, indicating stronger cohesion, and contained more pronouns. They found the MT system to be relatively successful at transferring these chains to the target language. For the news text, the characteristics of the referential chains in the output were similar to the statistics of human translations; for the literary text, there was a slight tendency towards underexpression of cohesive devices.

2.2.3. Improving lexical consistency and cohesion

15 There have been several attempts directly aimed at improving the consistency of lexical choice in the MT output. A multi-pass decoding approach to enforce consistent translation of recurring terms in a document was presented by Xiao et al. (2011) and was shown to improve the translation of English-Chinese newswire text. Their research was followed up by the work by Ture, Oard and Resnik (2012) cited above, which achieved improvements for Chinese-English and Arabic-English by designing features to guide the second-pass translation process instead of manipulating the phrase table as Xiao et al. (2011) did.

16 Alexandrescu and Kirchhoff (2009) described an approach based on graph-based learning to favor similar translations for similar input sentences by considering similarity both between training and test sentences and between pairs of test sentences, which led to large improvements for Italian-English and Arabic-English SMT tasks.

17 Lexical cohesion has been addressed in a somewhat different way with cache-based models, which adapt n-gram distributions or phrase translation probabilities by favoring events that have occurred in a certain context window that can span across sentence boundaries. Modest improvements with this approach have been demonstrated with a corpus of medical texts (Tiedemann, 2010a), while the same technique failed when applied to newswire text (Tiedemann, 2010b). One significant problem is that the cache easily gets contaminated with noise, and that it can contribute to the propagation of bad translations to the following sentences. More recently, improvements have been demonstrated with a more sophisticated caching technique that initializes the cache with statistics from similar documents found with information retrieval methods and keeps the noise level in check with the help of a topic model based on Latent Dirichlet Allocation (LDA) (Gong, Zhang & Zhou, 2011).

18 There have been a few attempts to use methods based on Latent Semantic Analysis (LSA) and LDA to achieve lexical cohesion under a topic model. Kim and Khudanpur (2004) used cross-lingual LSA to perform domain adaptation of language models in one language (assumed to suffer from scarce resources) given an adaptation corpus in another language. Zhao and Xing (2006) presented an approach to word alignment named BiTAM based on bilingual topic models, which they then extended to cover SMT decoding as well (Zhao & Xing, 2008). A similar technique based on a bilingual variant of LDA was used by Tam, Lane

and Schultz (2007) for adapting language models and phrase tables. Simpler and more recent approaches include the one by Gong, Zhang and Zhou (2010), who adapted SMT phrase tables with monolingual LDA, and Ruiz and Federico (2011), who implicitly trained bilingual LSA topic models by concatenating short pieces of text in both languages before training the model, and used these topic models for language model adaptation. Eidelman, Boyd-Graber and Resnik (2012) adapted features in the phrase table based on an LDA topic model. They compared adaptation at the sentence level with per-document adaptation and found that, while both approaches work, sentence-level adaptation gives marginally better results on their Chinese-English tasks.

2.3. Addressing discourse explicitly

19 Very recently, SMT researchers have begun to address more explicitly discourse-related problems. In this section, we give an overview of a number of efforts that reflect a more conscious decision to tackle the problems of document- or discourse-level translation. Recent research into improved translation of anaphoric pronouns, which would naturally fall under this heading as well, will be discussed in greater detail in Section 3.1.

2.3.1. Translating discourse connectives

20 The translation of discourse connectives is one challenge that has recently come into focus. In a corpus study, Cartoni et al. (2011) compared parts of the Europarl multilingual corpus (Koehn, 2005) that were originally written in French with other parts translated into French from English, German, Italian and Spanish. They found that the different subcorpora used fairly similar vocabulary in general, but that discourse connectives had significantly different distributions depending on the original source language of the text. They also noticed that it is fairly common for translators to introduce discourse connectives not explicitly found in the source language, and less common to leave out connectives present in the source. Later work from the same group (Meyer et al., 2011b) contrasted findings from a corpus study based on manual annotation with results obtained from the exploration of parallel corpora. Detailed results of the study are not contained in the published abstract.

21 Meyer et al. (2011a) and Meyer (2011) investigated automatic disambiguation of polysemous discourse connectives. They proposed a "translation spotting" annotation scheme for corpus data that marks up words that can be translated in different ways with their correct translation, which they call "transpot", instead of explicitly annotating linguistic features (Popescu-Belis et al., 2012). Disambiguating connectives with an automatic classifier before running a phrase-based SMT system resulted in small improvements in translation quality (Meyer, 2011; Meyer & Popescu-Belis, 2012; Meyer et al., 2012). Meyer et al. (2012) presented a family of automatic and semi-automatic evaluation scores to measure the accuracy of discourse connective translation in order to obtain a more meaningful assessment of progress on this problem than what a general-purpose measure like BLEU can deliver.

2.3.2. *Modeling verb tense*

22 Gong et al. (2012) presented a cross-sentence model to control the generation of the correct verb tenses in the MT output. This is a problem that occurs in the translation from Chinese to English because Chinese verbs are not morphologically marked for tense, whereas generating correct English output requires selecting the right tense form. They used n-gram-like features on the target side to model English verb tense sequence, with two different models to capture the sequence of verb tenses within a sentence and across sentences, respectively. Their cross-sentence model is just a sequence model over the tenses of the main verbs in each sentence. Sentences are processed in order, and information about the tense of the main verb generated is passed on to the following sentences so that the tense of the next verb can be conditioned on this information. By applying this model, they achieved an improvement of up to 0.8 BLEU points on a Chinese-English task.

2.3.3. *Algorithmic challenges*

23 One problem that recurs in different types of discourse-related SMT work is the difficulty of exploiting discourse-wide features because of the limitations of the decoding algorithm, which carries out the actual translation functions in an SMT system. Current systems almost universally use a variant of the dynamic programming beam search algorithm described by Koehn, Och and Marcu (2003) for decoding. This algorithm combines good search performance with high efficiency thanks to a dynamic programming technique exploiting the locality of the models, making it difficult or impossible to integrate models whose dependencies require considering a context larger than a window of five or six words. In past research, this problem was addressed mostly by handling cross-sentence dependencies in components outside the decoder, e.g., by decoding in two passes (Le Nagard & Koehn, 2010; Xiao et al., 2011; Ture, Oard & Resnik, 2012) or by using a special decoder driver module to annotate the decoder's input and retrieve the required information from its output (Hardmeier & Federico, 2010; Gong et al., 2012). More recently, Hardmeier, Nivre and Tiedemann (2012) presented a decoding algorithm based on local search that permits the inclusion of cross-sentence feature functions directly into the decoding process, opening up new ways to design discourse-wide models.

2.3.4. *Text-level MT evaluation*

24 A recurring issue in all discourse-related MT work is the problem of evaluation. The most popular automatic MT evaluation measure, BLEU (Papineni et al., 2002), calculates scores by measuring the overlap of low-order n-grams (usually up to 4-grams) between the output of the MT system and one or more reference translations. This score is insensitive to textual patterns that extend beyond the size of the n-grams, and it favors systems relying on strong n-gram models over other types of MT systems (Callison-Burch, Osborne & Koehn, 2006). It has been pointed out by various authors (Le Nagard & Koehn, 2010; Hardmeier & Federico, 2010; Guillou, 2011; Meyer et al., 2012) that this evaluation measure may not be adequate to

guide research on specific discourse-related problems, and more targeted evaluation scores have been devised for the translation of pronominal anaphora (Hardmeier & Federico, 2010) and discourse connectives (Meyer et al., 2012).

25 There has also been some effort to exploit discourse information to improve the evaluation of MT in general, independently of specific features in the MT systems tested. Comelles et al. (2010) proposed an MT evaluation metric based on Discourse Representation Theory (Kamp & Reyle, 1993), which takes into account features like coreference relations and discourse relations to assess the quality of MT output. Unfortunately, their metric does not have a higher correlation with human quality judgments than standard sentence-level MT evaluation metrics in the MetricsMATR shared task (Callison-Burch et al., 2010). However, it could be argued that the metric evaluation in the shared task itself was biased since the document-level human scores evaluated against were approximated by averaging human judgments of sentences seen out of context, so it is unclear to what extent the evaluation of a document-level score can be trusted.

26 Wong and Kit (2012) proposed extending sentence-level evaluation metrics, such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006) or METEOR (Banerjee & Lavie, 2005), with a component to measure lexical cohesion. For this, they used measures of word repetition in the text, after applying either just stemming or semantic relatedness according to similarity in WordNet (Fellbaum, 1998). They found that there was a positive correlation between their lexical cohesion scores and human quality judgments, and that they could improve the correlation of BLEU and TER, but not METEOR, by combining them with the cohesion scores. In finding a positive correlation between lexical cohesion as measured by word repetition in MT output and human quality judgments, their results seem to be inconsistent with those of Carpuat and Simard (2012) discussed above, a discrepancy that should be investigated further to pin down the role of lexical cohesion in MT quality.

2.3.5. Concluding remarks

27 To sum up, there is still little published work about discourse in SMT. Some work can be related to lexical cohesion even though most of it was written with domain adaptation or similar problems in mind. A number of researchers investigated WSD for SMT a few years ago. Their results underline the necessity of tight integration between the SMT system and the external discourse-related components. Explicit modeling of discourse phenomena for SMT has gained some attention recently with studies on pronominal anaphora and discourse connectives, but there are no strong results or proven methods available as yet.

3. Pronominal anaphora

28 After this overview of discourse-related SMT research in general, we shall now focus on the challenge of translating pronominal anaphora in SMT. Anaphora translation is one of the few discourse-level problems that have been studied by

different research groups, but it is far from being solved. Our analysis of the existing work on this topic and its shortcomings serves as an illustration of the difficulties inherent in discourse modeling for SMT. At the same time, we hope to make a contribution to the further progress of anaphora translation research by summing up and evaluating the existing approaches.

3.1. Existing work

29 Pronominal anaphora is the use of a pronoun to refer to an entity mentioned earlier in the discourse. This happens very frequently in most types of connected text. As an example, consider the following text passage, where *it* in the second sentence refers to *Poland* in the first:

[1] *Poland* is nevertheless pressing for observer status for non-euro members at euro zone meetings. History suggests *it* is unlikely to succeed.
(*The New York Times*, 10 Nov 2011)

30 When translating into French, *it* should be translated as *elle*, since *la Pologne* (Poland) has feminine gender. Had the name of the country in the first sentence been *le Portugal*, then the correct translation would have been the masculine *il*.

31 The usage and distribution of pronouns differ between languages (Russo et al., 2011). When an anaphoric pronoun is translated into a language with gender and number agreement, the correct form needs to be chosen according to the gender and number of the translation of its antecedent. Corpus studies have shown that this can be a problem for both statistical and rule-based MT systems, resulting in a potentially large number of mistranslated pronouns depending on language pair and text type (Hardmeier & Federico, 2010; Scherrer et al., 2011). Even though the translation and language models of SMT can manage to translate pronouns correctly in surprisingly many cases, the fact that the SMT system knows nothing about the meaning of pronouns can lead to pronouns being mistranslated in quite unexpected ways, such as in the following example (*newstest2009*[1]):

[2] *Input:* Elena first slaps Luca, then kisses him.
MT output: Elena première claques Luca, alors elle m'embrassait.
(*newstest2009*)

32 Le Nagard and Koehn (2010) approached the pronoun translation problem in phrase-based SMT by processing documents in two passes. The English input text was run through a coreference resolver developed by the authors *ad hoc*, and translation was performed with a regular SMT system to obtain French translations of the antecedent noun phrases (NPs). Then the anaphoric pronouns of the English text were annotated with the gender and number of the French translation of their

1. Examples marked *newstest2009* are selected from the 2009 test set of the shared tasks of the Workshop on Statistical Machine Translation (Callison-Burch et al., 2012).

antecedent and translated again with another MT system whose phrase tables had been annotated in the same way. This did not result in any noticeable increase in translation quality, which the authors put down to the insufficient quality of their coreference resolution system. However, in a later application of the same approach to an English-Czech system, no clearly positive results were obtained despite the use of data manually annotated for coreference (Guillou, 2011 and 2012).

33 Hardmeier and Federico (2010) took on the same challenge with a one-pass system that directly incorporates the processing of coreference links into the decoding step. Pronoun coreference links were annotated with the BART software (Broscheit et al., 2010). The authors then added an extra feature to the decoder to model the probability of a pronoun given its antecedent. Sentence-internal coreference links were handled completely within the SMT dynamic programming algorithm. For links across sentence boundaries, the translation of the antecedent was extracted from the MT output after translating the sentence containing it, and it was held fixed when the referring pronoun was translated. In this work, no improvement in BLEU score (Papineni et al., 2002), the most popular MT evaluation metric, was achieved for English-German translation, but a slight improvement was found with an evaluation metric targeted specifically to pronoun coreference. A subsequent attempt to apply the same technique to the language pair English-French was largely unsuccessful (Hardmeier et al., 2011).

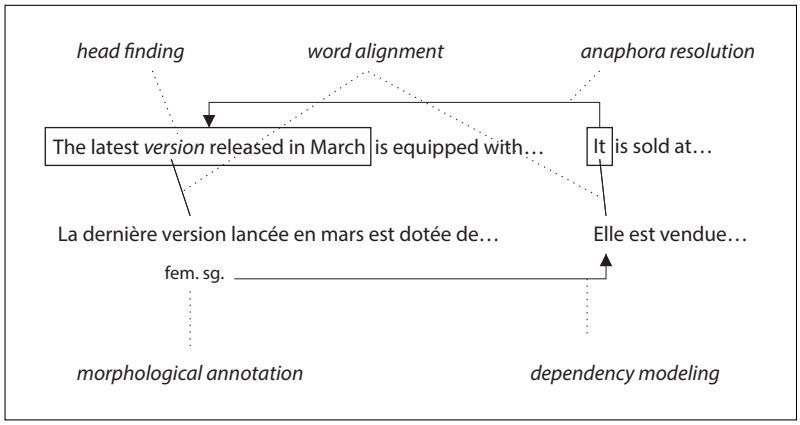

Figure 1. Components of an anaphora handling model for SMT

34 Figure 1 illustrates the different parts of the pronoun translation task as it was implemented by Hardmeier and Federico (2010). It shows the translation of an English input text into French. In order to translate the anaphoric pronoun *it*, its antecedent is determined with an *anaphora resolution* system, which resolves it to the NP "the latest version released in March". A *head finder* identifies the head word of this NP, the word "version", whose translation is assumed to carry the required gender and number information. Using *word alignments*, the translation of this word in the target language with its *morphological annotation* is looked up.

The morphological features are then used as input to a *word dependency model* that predicts the pronoun to be used. The approach embraced by Le Nagard and Koehn (2010) and Guillou (2011 and 2012) is very similar, but does not use an explicit word dependency model. Instead, a similar effect is achieved by a second decoding pass with an annotated phrase table.

35 A somewhat different problem was addressed by Russo, Loáiciga and Gulati (2012a and b). They considered the generation of subject pronouns when translating from pro-drop languages into languages that require pronominal subjects to be realized explicitly, conducting a corpus study and examining the output of a rule-based and a statistical MT system. Their work focused on identifying where to insert pronouns with the help of rule-based preprocessing and statistical postprocessing; they made no attempt to resolve anaphoric references and resorted to inserting majority class (masculine) pronouns whenever there was an ambiguity.

36 In a thesis proposal paper submitted to the Week of Doctoral Students, Novák (2011) discussed a number of ways to make use of anaphora resolution for SMT, specifically from English into Czech using the deep syntactic TectoMT system developed at Charles University in Prague. He presented an analysis of errors made by the MT system and found that about half of the occurrences of the pronoun *it* in his corpus were expletives or referred anaphorically to non-NP constituents. In these cases, the obvious translation of *it* with a Czech neuter pronoun would most often be correct. The pronoun was also consistently translated with a Czech neuter when it did have NP reference, and a substantial part of these cases were wrong.

3.2. Challenges in anaphora translation

37 Despite the efforts several researchers have undertaken to improve the translation of anaphoric pronouns with SMT, no convincing and reproducible positive results have been published as yet. Guillou (2012) discussed a number of reasons for the disappointing performance of SMT systems with anaphora handling. In particular, she identified four main sources of error:

- identification of anaphoric vs. non-anaphoric pronouns;
- anaphora resolution;
- identification of the head of the antecedent NPs, from which gender and number features are extracted;
- word and phrase alignment between source and target text.

38 We agree with Guillou that these items are important factors that must be addressed in future work, but we believe that the difficulties in demonstrating an improvement of SMT systems by explicit pronoun handling are due to an even wider range of accumulating deficiencies in various components of the experimental setup. We have identified six principal factors that present risks to pronoun-aware SMT systems and may help to explain the failure of existing research to find solutions. The sources of

error listed by Guillou (2012) can be subsumed under these headings; however, our analysis is considerably broader and discusses some points that have been taken for granted in most previous research. In this section, we examine these challenges in some more detail, starting with risks external to the pronoun translation approaches proper before moving towards deficiencies inherent in the methods that were tested.

3.2.1. Insufficient performance of baseline SMT system

39 Models for anaphoric pronouns target a very specific linguistic phenomenon by affecting few words in the output text. This can only be successful if the translation as a whole is reasonably good; no pronoun translation model will be able to achieve significant improvements if what the underlying SMT system outputs without its help is mostly gibberish. It is well known that some language pairs are much more difficult for SMT than others, for instance because of word order differences or complex target language morphology. In other cases, out-of-vocabulary words in the input text may make the translation unreliable. When this happens, there is not much that a pronoun model can do to improve the translation because it is too specifically focused on a single phenomenon.

40 Insufficient baseline performance was mentioned by Hardmeier and Federico (2010) as a major problem for their English-German system. A similar hint is made by Guillou (2011: 49), who notes that "[o]ne of the major difficulties that [human evaluators] encountered during the evaluation was in connection with evaluating the translation of pronouns in sentences which exhibit poor syntactic structure". This suggests that, at least in some cases, the translations output by her English-Czech MT system were so poor as to render pronoun-specific evaluation essentially meaningless.

41 By contrast, the output of state-of-the-art English-French SMT systems is to a large extent intelligible if not perfect. It sometimes happens that the SMT system garbles the syntax of a sentence, such as in the following examples:

[3] *Input:* We don't have stewardesses, we've been against it from the very beginning.
MT output: Nous n'avons pas, nous avons été hôtesses contre elle dès le début.
(*newstest2009*)

[4] *Input:* And this time, Hurston's old neighbors saw her as a savior.
MT output: Et cette fois, l'ancienne Hurston voisins a vu son comme un sauveur.
(*newstest2009*)

42 These cases are fairly rare, however, and it is reasonable to assume that this was the case also for the anaphora-sensitive English-French systems described in the literature (Le Nagard & Koehn, 2010; Hardmeier et al., 2011). Generally, there is little that researchers interested in anaphora can do about this problem except working on an easier language pair while waiting for the progress of SMT research in general. English-French is probably a good choice in this respect.

3.2.2. *Insufficient performance of coreference resolution system*

43 The performance of any MT system that attempts to model pronominal anaphora explicitly is dependent on the quality of its anaphora resolution component. When many anaphoric links are resolved incorrectly, the model may degrade performance on average rather than improve it. To see why, consider that an SMT system with no explicit anaphora handling component will not emit pronouns randomly; rather, the system is likely to have a preference for the pronouns that are most frequent in the training corpus. If the test set is homogeneous with the training data, this may very well be the correct choice in many cases. An example of this can be seen in the pronoun translation corpus study conducted by Hardmeier and Federico (2010). Their SMT system has a strong preference for translating the ambiguous German pronoun *sie* as *they* or *them* rather than as *she* or *her*. As a result, pronoun translation errors are very frequent in documents featuring a female protagonist, whereas many other documents are hardly affected. Overall, anaphora resolution is a difficult task in itself, and inadequate performance of the coreference resolver has been advanced as an explanation for disappointing experimental results at least by Le Nagard and Koehn (2010).

44 Pronouns are notoriously difficult for anaphora resolution systems to resolve correctly when they do not refer to a NP. On the one hand, this applies to expletive pronouns such as *it* in *it is raining*, which are not used anaphorically at all. Detecting expletives is a rather difficult problem. Le Nagard and Koehn (2010) implemented a rule-based system for this task (Paice & Husk, 1987), which performed surprisingly well for them at a precision and recall of 83%; however, the same system has been shown to perform considerably worse on different corpus data (Evans, 2001). Currently one of the best expletive detectors for English is the one by Bergsma and Yarowsky (2011), which achieves high accuracy on a variety of test sets. For the French pronoun *il*, there is a publicly available rule-based system that is reported to achieve an accuracy of more than 96% (Danlos, 2005).

45 Low recall for expletive classification means that a substantial part of the expletive pronouns in a text will be incorrectly linked to an antecedent. As an example, consider the following two sentences, where the automatic coreference resolution system used by Hardmeier et al. (2011) incorrectly chose to link the non-referring pronoun *It* in the second sentence to the word *it* in the first and to create a coreference chain *price – it – It*:

[5] Napi's basket suggested that this latter was a near impossibility, since we found that the price was up by just a shade over 10 percent on last year's quite high base *price*, even where *it* was most expensive. *It* does appear, though, that flour suppliers are in a stronger position than egg producers, for they have managed to force their drastic price increases onto the multinationals.
(*newstest2009*)

46 On the other hand, pronouns may refer to events rather than entities expressed by NPs, as in the following example:

[6] *Input:* He made a scandal out of *it* when the Prefecture ordered the dissolution of the municipal council.
MT output: Il fait un scandale de la préfecture lorsqu'elle a ordonné la dissolution du conseil municipal.
(*newstest2009*)

47 This type of coreference is somewhat neglected by current coreference resolution systems (Pradhan et al., 2011), so pronouns with event anaphora will often be resolved incorrectly as referring to a NP. At the same time, both expletives and event anaphora may be relatively easy for a naive SMT system to get right, since they are generally rendered with a small set of common pronouns such as *it* in English or *il*, *ça*, *cela* in French. In these cases, incorrect anaphora resolution greatly increases the risk of mistranslation.

48 One consideration that previous research has spent little thought on is the output format of the anaphora resolver. Usually, coreference resolution systems output coreference chains, disjoint sets of those mentions in a document that refer to the same entity. There are different ways of decomposing the task of identifying coreference chains into elementary classification or ranking decisions (Ng, 2010). This is not what has been used by the existing pronoun-enhanced SMT systems, however. Both Le Nagard and Koehn (2010) and Hardmeier and Federico (2010) used pairwise links between pronouns and their direct antecedents rather than coreference chains as input for their systems. Moreover, both systems rely on the one-best output of their anaphora resolvers. In this way, the SMT systems are completely dependent on single, local decisions of an anaphora resolution system that is known to be unreliable. Using coreference chains or allowing the anaphora resolution system to output multiple competing antecedent candidates might improve the robustness of the system as a whole.

3.2.3. Translation divergences

49 All the anaphora-enabled SMT systems discussed above have made the tacit assumption that anaphoric pronouns should be translated with anaphoric pronouns in the vast majority of cases. While it is usually acknowledged that this might not always be true, the cases in which it is not have been regarded as rare exceptions, possibly due to errors in the anaphora resolution process. A very brief and fairly superficial experiment with some corpus data sheds some doubt on this assumption.

50 For the experiment, we took two publicly available English-French parallel corpora published in connection with recent MT shared tasks: the *news-commentary* corpus is a parallel corpus of news texts of around 3 million words from the 2011 Workshop on SMT shared task (Callison-Burch et al., 2011). The *TED* corpus is the training corpus of the WIT[3] corpus collection (Cettolo, Girardi & Federico, 2012). It consists of around 2.8 million words of transcribed talks from the TED (Technology, Entertainment and Design) conferences. For these two corpora, we computed automatic word alignments with a standard training procedure for

phrase-based SMT (Koehn, 2010) and annotated the French part of the corpus with part-of-speech tags using the *TreeTagger* software (Schmid, 2003). We then counted how often the English pronouns *he*, *she*, *it* and *they* were aligned to a word tagged as any type of pronoun or determiner in French. Determiners were included to account for the fact that the French direct object pronouns *le*, *la* and *les* are frequently mistagged as definite articles. The following table shows the percentage of the English pronouns that were aligned to a French pronoun counterpart:

	news-commentary	*TED*
Sample size	25,474	71,426
he	81.8%	83.2%
she	81.4%	84.7%
it	75.4%	72.7%
they	81.8%	83.6%
Total	78.2%	77.1%

Table 1. Percentage of pronouns aligned to pronouns in the reference translation

51 In both genres, we found that more than 20% of the occurrences of the four pronouns examined, including between 15 and 20% of the tokens *he*, *she* and *they*, which cannot be used as expletives, were not aligned to a pronoun in the other language. Note that the percentages are very similar in both corpora, even though the slightly smaller *TED* corpus contains almost three times as many pronouns as the *news-commentary* corpus. Some part of the pronouns that are not aligned to pronouns may be due to alignment errors or very free translations in the parallel corpus, but it is not difficult to find examples of pronouns that are quite legitimately translated as non-pronouns even in fairly literal translation. In some cases, it is almost impossible to use a pronoun in the target language while still retaining fluency (*TED-dev2010*[2]):

[7] *Input:* Initially, all we did was *autograph* ***it***.
Reference: Pour commencer, nous avons juste *mis notre autographe.*
Gloss: To begin, we just put our autograph.
(*TED-dev2010*)

[8] *Input: Most of* ***them*** are ordinary digital camera photos.
Reference: La plupart sont des photos d'appareils numériques ordinaires.
Gloss: The majority are ordinary digital camera photos.
(*TED-dev2010*)

2. Examples marked *TED-dev2010* are selected from the *dev2010* test set of the WIT[3] corpus (Cettolo, Girardi & Federico, 2012).

52 In other cases, the missing pronouns are due to slight changes in wording in the reference translation. Even in the following examples, however, we would still consider the translation to be quite close:

[9] *Input:* **It** doesn't create the distortion of reality; **it** creates the dissolution of reality.
Reference: **Elle** ne provoque pas une déformation de la réalité; *mais plutôt* une dissolution de la réalité.
Gloss: **It** doesn't create a distortion of reality; *but rather* a dissolution of reality.
(*TED-dev2010*)

[10] *Input:* But *the thing about tryptamines is* ***they*** *cannot* be taken orally *because* ***they****'re denatured* by an enzyme found naturally in the human gut.
Reference: Par contre *les tryptamines ne peuvent pas* être consommées par voie orale *étant dénaturé*[*e*]*s* par une enzyme se trouvant de façon naturelle dans l'intestin de l'homme.
Gloss: However *tryptamines cannot* be consumed orally *being denatured* by an enzyme found naturally in the human gut.
(*TED-dev2010*)

53 The composition of this residual set of pronouns not aligned to pronouns is a matter that deserves further investigation. Depending on whether the lack of correspondence between the source and the target language is a result of processing error or a natural feature of the data, it may be beneficial to handle these pronouns specially, or remove them, during system training, or to account for them explicitly in the models used at decoding time.

3.2.4. Inadequate evaluation

54 It is widely recognized that automatic evaluation of pronoun translation is difficult and that existing methods are unreliable (Hardmeier & Federico, 2010; Le Nagard & Koehn, 2010; Guillou, 2011). Popular MT evaluation metrics such as BLEU (Papineni et al., 2002) score the MT output by comparing it to one or more reference translations. This approach is fraught with problems. Since it is completely unspecific and assigns the same weight to any overlap with the reference, it is not particularly sensitive to the type of improvements targeted by a pronoun translation component, which affect only a few words in a text.

55 Hardmeier and Federico (2010) addressed this shortcoming by using a precision/recall-based measure counting the overlap of pronoun translations between the MT output and a reference translation. While increasing the sensitivity to pronoun changes, this measure retains another serious drawback of a reference-based pronoun evaluation in that it judges correctness by comparing the translation of a pronoun in the MT output with the translation found in a reference translation and assumes that they should be the same. However, this assumption is flawed: it does not necessarily hold if the MT system selects a different translation for the antecedent

of the pronoun. If this is the case, the only meaningful way to check the correctness of a pronoun is by finding out whether it agrees with the antecedent selected by the system, even if the translation of the antecedent may be incorrect. As Guillou (2011) remarks, the accuracy of an evaluation method that checks pronouns against a reference translation also depends on the number of inflectional forms for pronouns in the target language. If pronouns are inflected for a number of different features in a given language, the probability of matching a pronoun exactly with a noisy system is very low, and it becomes difficult to measure progress before perfection is achieved.

56 More relevant conclusions about the quality of pronoun translation could be drawn by examining how the MT output renders the coreference chains found in the input and checking the pronouns referring to the same entity for consistency. The main difficulty here is that this makes the evaluation dependent on coreference annotations for the source language, leading to unreliable evaluation results when there are errors in the annotation. This evaluation strategy was adopted by Guillou (2011) and worked well for her since she had gold-standard coreference annotations for her test set. In the absence of gold-standard annotations, reliable automatic evaluation of pronoun translations seems difficult or impossible.

3.2.5. Error propagation

57 Both SMT pronoun models described in the literature (Le Nagard & Koehn, 2010; Hardmeier & Federico, 2010) model coreference as links between pairs of words, a referent (pronoun) and an antecedent. Longer coreference chains are decomposed into links between pairs of words. As a result, the pronoun models only consider information about the immediately preceding antecedent when handling a pronoun. This is particularly relevant if a coreference chain consists of a sequence of pronouns. If the SMT system, triggered by some other factor such as the n-gram model, mistranslates one of the pronouns in the chain, this error can easily be propagated to all later elements of the chain. This problem could be addressed either by processing the coreference links so that links pointing to an antecedent that is a pronoun are transitively extended until a full word form is reached, or by jointly optimizing over all coreference links in the document at the same time. The latter approach is incompatible with the dynamic programming decoding technique that is currently the most popular in SMT and requires a decoder capable of handling cross-sentence dependencies such as the one by Hardmeier, Nivre and Tiedemann (2012).

3.2.6. Model deficiencies

58 Le Nagard and Koehn (2010: 259) claim that "[their] method works in principle", if it was not for the poor performance of the coreference resolution system, and Hardmeier and Federico (2010) report minor improvements for the pronoun *it* in a pronoun-specific automatic evaluation with their method. Nevertheless, by demonstrating that performance remains unconvincing even when using gold-standard

coreference annotations (Guillou, 2011) and that the small improvements that have been achieved do not carry over to another language pair (Hardmeier et al., 2011), later work suggests that both methods are in need of refinement before they can deliver consistently useful results.

59 An interesting observation made by both Guillou (2011) and Hardmeier et al. (2011) is that SMT systems with explicit pronoun handling tend to generate more pronouns than required. The reason for this need not be the same for the two systems. In particular, in the English-Czech system, one difference between the languages is that Czech, unlike English, allows subject pronouns to be left out when the subject can be inferred from the context. The observed overgeneration effect may result from a reduced tendency of the second-pass system with its more focused pronoun translation distributions to drop pronouns, word removal being an event not explicitly accounted for in the standard phrase-based SMT model.

60 In the experiments by Hardmeier et al. (2011), anaphoric links were modeled by a bigram language model predicting pronouns given gender and number of the antecedent. The vocabulary of the predicted words was restricted to pronominal forms. Other words were treated as "out of vocabulary" by the model and penalized harshly. This may lead to a strong preference for translating every single pronoun as a pronoun, even when this is not an adequate translation, e.g., when the coreference system mistakenly resolved a non-referential pronoun.

61 Another potential problem is the source of the agreement features that are used for a particular antecedent. Existing work has used some sort of head finding algorithm to identify the syntactic head of the antecedent NP, looked up the corresponding target language words with the help of the word alignments and extracted morphological features from these target language words (Hardmeier & Federico, 2010; Guillou, 2011). As Guillou (2012) points out, both head finding and word alignment are prone to errors when done automatically; the same can be said for the morphological annotation that provides gender and number features. Both Hardmeier and Federico (2010) and Le Nagard and Koehn (2010) used a lexicon lookup based on the Le*fff* full form lexicon (Sagot et al., 2006) to determine the gender and number of French words; words not listed in this lexicon, such as some proper names, were handled heuristically if at all. Guillou (2011 and 2012) was able to circumvent some of these problems by using gold-standard annotations.

62 In sum, the existing pronoun models for SMT are clearly less than perfect, and pronoun overgeneration is a problem that has been observed repeatedly with different models. To improve the models, the reasons for this behavior should be examined more closely. It may be necessary to design an explicit model for dropping pronouns or translating them with non-pronouns. As pointed out earlier, research on anaphora resolution has had a tendency towards focusing on the prototypical case of anaphora with a nominal antecedent, but non-referential pronouns and event anaphora pose harder challenges to current systems. The same preference for prototypical problem instances can be observed in research on SMT pronoun

models; in SMT, however, the less frequent, non-prototypical cases may in fact be easier to handle for a naive system since, at least for target languages like French or German, agreement patterns are much less complex than for nominal antecedents. Consequently, there is a substantial risk of degrading performance by adding a pronoun model that mishandles these very categories.

3.3. Prospects for future research

63 In the preceding section, we gave an overview of possible reasons for the failure of past research efforts to improve the translation of pronominal anaphora with SMT. Let us now recapitulate these factors and discuss their relevance for SMT research in the near future.

64 The first factor we mentioned is baseline performance, which means the performance of all components of the SMT system except the ones we are interested in. What we can do here is select our baseline system so as to maximise the effect of the model we want to test. For pronoun translation, it seems important to choose a language pair with very good SMT performance as it is almost impossible to improve on an underperforming MT system with a pronoun model. At the same time, it is important that there be an interesting difference in pronoun systems between the source and the target language. From among the language pairs commonly used in SMT research, combinations of English with French or Spanish are probably good choices, whereas it may be more difficult to demonstrate improvements for language pairs like English-German or English-Czech.

65 Given the difficulty of the coreference resolution task, it is certainly important to use the best coreference resolution systems available. In order to discern the effects of imperfect coreference resolution from deficiencies of the pronoun models themselves, experiments with manually annotated coreference data such as those conducted by Guillou (2011) are very valuable, and the creation of parallel test sets with gold-standard coreference annotation for carefully chosen language pairs would be extremely useful. The availability of manually checked coreference annotations would also make it much easier to devise a reliable evaluation method for pronoun translation.

66 When it comes to the intrinsic weaknesses of the pronoun translation models, we believe that it will be necessary to review the models that have been proposed and to examine more closely the circumstances under which the existing pronoun models degrade the baseline performance. It will probably be necessary to find ways to control the overgeneration of pronouns. The impact of error propagation should be evaluated quantitatively and the problem addressed if it turns out to be necessary. Quite generally, considering the noise present in every step of a pronoun-aware SMT system and the number of steps involved, it may be necessary to design models that explicitly deal with the uncertainty present in each step, globally minimising the risk of choosing the wrong pronoun rather than putting together the one-best outputs of a large number of inaccurate components.

4. Conclusion

67 In this article, we have presented an overview and survey of the existing literature dealing with discourse-level phenomena in the context of SMT. It turns out that, even though explicit attempts to address discourse in SMT have been rare, there is a surprisingly large body of literature dealing with related problems from different perspectives such as "domain adaptation", "consistency enforcement" or "disambiguation". More focused interest in discourse proper has wakened very recently and has led to still ongoing research on phenomena such as pronominal anaphora and discourse connectives. We then went on to discuss the problem of pronoun generation in SMT, providing a more detailed survey of recent research activities and analyzing possible reasons for past failures in this task, which has proven extremely resistant to the initial approaches taken by different researchers.

68 Even though it seems obvious to the human language user that discourse-level information must be useful for translation, exploiting it successfully has turned out to be fairly difficult. Coupling SMT systems with existing modules such as WSD systems or anaphora resolvers often leads to negative results, unless those modules are specifically adapted to the task at hand. Low accuracy is a recurring problem; frequently, the noise of different system components will add up and cancel out almost all useful information. We believe that it is most promising to use the expertise accumulated in different fields of natural language processing to create systems specifically designed to be used together with SMT rather than relying on standard task definitions. This has been done with some success for WSD. Furthermore, it seems advisable to devise explicit ways of handling inaccuracies, e.g., by using probabilistic confidence measures, rather than putting blind trust in the output of intermediate system components and propagating early errors through long processing pipelines.

69 Undoubtedly, research about discourse in SMT stands at its beginning, and there are many formidable challenges to overcome. At the moment, independence between sentences is an assumption taken for granted by many researchers. We believe that a wide range of fascinating and important research problems becomes accessible once we overcome this limitation and start developing document-wide models and approaches.

Acknowledgments

70 We gratefully acknowledge the help of Jörg Tiedemann and Joakim Nivre in preparing the final version of this article.

References

Alexandrescu, A. & Kirchhoff, K. 2009. Graph-Based Learning for Statistical Machine Translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 119-127. Available online: http://aclweb.org/anthology-new/N/N09/N09-1014.pdf.

Banerjee, S. & Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Stroudsburg: Association for Computational Linguistics: 65-72. Available online: http://aclweb.org/anthology-new/W/W05/W05-0909.pdf.

Bergsma, S. & Yarowsky, D. 2011. NADA: A Robust System for Non-Referential Pronoun Detection. In I. Hendrickx et al. (ed.), *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, Faro, Portugal, October 6-7, 2011*. Berlin: Springer: 12-23.

Broscheit, S. et al. 2010. BART: A Multilingual Anaphora Resolution System. In K. Erk & C. Strapparava (eds.), *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*. Stroudsburg: Association for Computational Linguistics: 104-107. Available online: http://aclweb.org/anthology-new/S/S10/S10-1021.pdf.

Brown, P.F. et al. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16 (2): 79-85.

Brown, P.F. et al. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19 (2): 263-311.

Callison-Burch, C. et al. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 17-53. Available online: http://aclweb.org/anthology-new/W/W10/W10-1703.pdf.

Callison-Burch, C. et al. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 22-64. Available online: http://aclweb.org/anthology-new/W/W11/W11-2103.pdf.

Callison-Burch, C. et al. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 10-51. Available online: http://www.aclweb.org/anthology/W/W12/W12-3102.pdf.

Callison-Burch, C., Osborne, M. & Koehn, P. 2006. Re-Evaluating the Role of BLEU in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*. Stroudsburg: Association for Computational Linguistics: 249-256. Available online: http://aclweb.org/anthology-new/E/E06/E06-1032.pdf.

Carpuat, M. 2009. One Translation per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*. Stroudsburg: Association for Computational Linguistics: 19-27. Available online: http://aclweb.org/anthology-new/W/W09/W09-2404.pdf.

Carpuat, M. & Simard, M. 2012. The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 442-449. Available online: http://aclweb.org/anthology-new/W/W12/W12-3156.pdf.

Carpuat, M. & Wu, D. 2005. Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 387-394. Available online: http://aclweb.org/anthology-new/P/P05/P05-1048.pdf.

Carpuat, M. & Wu, D. 2007a. How Phrase Sense Disambiguation Outperforms Word Sense Disambiguation for Statistical Machine Translation. In A. Way & B. Gawronska (eds.), *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI 2007*. Skövde: University of Skövde, School of Humanities and Informatics: 43-52. Available online: http://mt-archive.info/TMI-2007-complete.pdf.

Carpuat, M. & Wu, D. 2007b. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics: 61-72. Available online: http://aclweb.org/anthology-new/D/D07/D07-1007.pdf.

Cartoni, B. et al. 2011. How Comparable Are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Stroudsburg: Association for Computational Linguistics: 78-86. Available online: http://aclweb.org/anthology-new/W/W11/W11-1211.pdf.

Cettolo, M., Girardi, C. & Federico, M. 2012. Wit^3: Web Inventory of Transcribed and Translated Talks. In M. Cettolo et al. (eds.), *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*. Trento: Fondazione Bruno Kessler: 261-268. Available online: http://mt-archive.info/EAMT-2012-complete.pdf.

Chan, Y.S., Ng, H.T. & Chiang, D. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 33-40. Available online: http://aclweb.org/anthology-new/P/P07/P07-1005.pdf.

Chiang, D. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics* 33 (2): 201-228.

Comelles, E. et al. 2010. Document-Level Automatic MT Evaluation Based on Discourse Representations. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 333-338. Available online: http://aclweb.org/anthology-new/W/W10/W10-1750.pdf.

Danlos, L. 2005. Automatic Recognition of French Expletive Pronoun Occurrences. In *IJCNLP-05. Second International Joint Conference on Natural Language Processing. Companion Volume to the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts*. Asian Federation of Natural Language Processing: 73-78. Available online: http://aclweb.org/anthology-new/I/I05/I05-2013.pdf.

Eidelman, V., Boyd-Graber, J. & Resnik, P. 2012. Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Stroudsburg: Association for Computational Linguistics: 115-119. Available online: http://aclweb.org/anthology-new/P/P12/P12-2023.pdf.

Evans, R. 2001. Applying Machine Learning Toward an Automatic Classification of it. *Literary and Linguistic Computing* 16 (1): 45-57.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge (Mass.): MIT Press.

Forcada, M.L. et al. 2011. Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation. *Machine Translation* 25 (2): 127-144.

Gale, W.A., Church, K.W. & Yarowsky, D. 1992. One Sense per Discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. Burlington: M. Kaufmann: 233-237.

Giménez, J. & Màrquez, L. 2007. Context-Aware Discriminative Phrase Selection for Statistical Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 159-166. Available online: http://aclweb.org/anthology-new/W/W07/W07-0719.pdf.

Gong, Z. et al. 2012. N-Gram-Based Tense Models for Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics: 276-285. Available online: http://aclweb.org/anthology-new/D/D12/D12-1026.pdf.

Gong, Z., Zhang, M. & Zhou, G. 2010. Statistical Machine Translation Based on LDA. In *2010 4th International Universal Communication Symposium (IUCS 2010)*. Institute of Electrical and Electronics Engineers (IEEE): 286-290.

Gong, Z., Zhang, M. & Zhou, G. 2011. Cache-Based Document-Level Statistical Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics: 909-919. Available online: http://aclweb.org/anthology-new/D/D11/D11-1084.pdf.

Guillou, L. 2011. *Improving Pronoun Translation for Statistical Machine Translation (SMT)*. Master's thesis. University of Edinburgh, School of Informatics.

Guillou, L. 2012. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 1-10. Available online: http://aclweb.org/anthology-new/E/E12/E12-3001.pdf.

Hardmeier, C. et al. 2011. The Uppsala-FBK Systems at WMT 2011. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 372-378. Available online: http://aclweb.org/anthology-new/W/W11/W11-2144.pdf.

Hardmeier, C. & Federico, M. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT)*. 283-289. Available online: http://uu.diva-portal.org/smash/get/diva2:420761/FULLTEXT01.

Hardmeier, C., Nivre, J. & Tiedemann, J. 2012. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics: 1179-1190. Available online: http://aclweb.org/anthology-new/D/D12/D12-1108.pdf.

Hatim, B. & Mason, I. 1990. *Discourse and the Translator*. Language in social life series. London – New York: Longman.

Isabelle, P. & Foster, G. 2006. Machine Translation: Overview. In E.K. Brown et al. (eds.), *Encyclopedia of Language and Linguistics*. Amsterdam – Boston – Heidelberg: Elsevier. Vol. 7: 404-422.

Kamp, H. & Reyle, U. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht – Boston – London: Kluwer Academic Publishers.

Kim, W. & Khudanpur, S. 2004. Cross-Lingual Latent Semantic Analysis for Language Modeling. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Piscataway (N.J.): Institute of Electrical and Electronics Engineers (IEEE). Vol. 1: 257-260.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*. 79-86. Available online: http://www.mt-archive.info/MTS-2005-Koehn.pdf.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge – New York: Cambridge University Press.

Koehn, P., Och, F.J. & Marcu, D. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 48-54. Available online: http://aclweb.org/anthology-new/N/N03/N03-1017.pdf.

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 MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 252-261. Available online: http://aclweb.org/anthology-new/W/W10/W10-1737.pdf.

Marcu, D., Carlson, L. & Watanabe, M. 2000. The Automatic Translation of Discourse Structures. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 9-17. Available online: http://aclweb.org/anthology-new/A/A00/A00-2002.pdf.

Meyer, T. 2011. Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation. In *Proceedings of the ACL 2011 Student Session*. Stroudsburg: Association for Computational Linguistics: 46-51. Available online: http://aclweb.org/anthology-new/P/P11/P11-3009.pdf.

Meyer, T. et al. 2011a. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of the SIGDIAL 2011 Conference*. Stroudsburg: Association for Computational Linguistics: 194-203. Available online: http://aclweb.org/anthology-new/W/W11/W11-2022.pdf.

Meyer, T. et al. 2011b. Disambiguating Discourse Connectives Using Parallel Corpora: Senses vs. Translations. In *Proceedings of the Corpus Linguistics 2011 Conference (Birmingham, 20-22 July 2011)*.

Meyer, T. et al. 2012. Machine Translation of Labeled Discourse Connectives. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*. The Association for Machine Translation in the Americas (AMTA). Available online: http://amta2012.amtaweb.org/AMTA2012Files/papers/119.pdf.

Meyer, T. & Popescu-Belis, A. 2012. Using Sense-Labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*. Stroudsburg: Association for Computational Linguistics: 129-138. Available online: http://aclweb.org/anthology-new/W/W12/W12-0117.pdf.

Mitkov, R. 1999. Introduction: Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP. *Machine Translation* 14 (3-4): 159-161.

Ng, V. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 1396-1411. Available online: http://aclweb.org/anthology-new/P/P10/P10-1142.pdf.

Novák, M. 2011. Utilization of Anaphora in Machine Translation. In J. Safrankova & J. Pavlu (eds.), *WDS 2011 – Proceedings of Contributed Papers: Part I – Mathematics and Computer Sciences*. Prague: Matfyzpress: 155-160. Available online: http://ufal.mff.cuni.cz/~mnovak/papers/WDS_2011_paper.pdf.

Paice, C.D. & Husk, G.D. 1987. Towards the Automatic Recognition of Anaphoric Features in English Text: The Impersonal Pronoun "it". *Computer Speech and Language* 2 (2): 109-132.

Papineni, K. et al. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 311-318. Available online: http://aclweb.org/anthology-new/P/P02/P02-1040.pdf.

Popescu-Belis, A. et al. 2012. Discourse-Level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Stroudsburg: Association for Computational Linguistics: 2716-2720. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/255_Paper.pdf.

Pradhan, S. et al. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*. Stroudsburg: Association for Computational Linguistics: 1-27. Available online: http://aclweb.org/anthology-new/W/W11/W11-1901.pdf.

Ruiz, N. & Federico, M. 2011. Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 294-302. Available online: http://aclweb.org/anthology-new/W/W11/W11-2133.pdf.

Russo, L. et al. 2011. Étude inter-langues de la distribution et des ambiguïtés syntaxiques des pronoms. In M. Lafourcade & V. Prince (eds.), *TALN 2011/RECITAL 2011 (18e conférence annuelle sur le Traitement automatique des langues naturelles, 27 juin-1er juillet 2011, Montpellier)*. Montpellier: AVL Diffusion. Vol. 2: 279-284.

Russo, L., Loáiciga, S. & Gulati, A. 2012a. Improving Machine Translation of Null Subjects in Italian and Spanish. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 81-89. Available online: http://aclweb.org/anthology-new/E/E12/E12-3010.pdf.

Russo, L., Loáiciga, S. & Gulati, A. 2012b. Italian and Spanish Null Subjects. A Case Study Evaluation in an MT Perspective. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Stroudsburg: Association for Computational Linguistics: 1779-1784. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/813_Paper.pdf.

Sagot, B. et al. 2006. The Lefff 2 Syntactic Lexicon for French: Architecture, Acquisition, Use. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. Stroudsburg: Association for Computational Linguistics: 1348-1351. Available online: http://www.lrec-conf.org/proceedings/lrec2006/pdf/810_pdf.pdf.

Sanders, T. & Pander Maat, H. 2006. Cohesion and Coherence: Linguistic Approaches. In E.K. Brown et al. (eds.), *Encyclopedia of Language and Linguistics*. Amsterdam – Boston – Heidelberg: Elsevier. Vol. 2: 591-595.

Scherrer, Y. et al. 2011. La traduction automatique des pronoms. Problèmes et perspectives. In M. Lafourcade & V. Prince (eds.), *TALN 2011/RECITAL 2011 (18e conférence annuelle sur le Traitement automatique des langues naturelles, 27 juin-1er juillet 2011, Montpellier)*. Montpellier: AVL Diffusion. Vol. 2: 185-190.

Schmid, H. 2003. Probabilistic Part-of-Speech Tagging Using Decision Trees. In H.L. Somers & D.B. Jones (eds.), *New Methods in Language Processing*. London: Routledge: 154-164.

Snover, M. et al. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas – "Visions for the Future of Machine Translation"*. The Association for Machine Translation in the Americas (AMTA): 223-231. Available online: http://www.mt-archive.info/AMTA-2006-Snover.pdf.

Specia, L., Sankaran, B. & das Graças Volpe Nunes, M. 2008. N-best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing: 9th International Conference, CICLing 2008, Haifa, Israel, February 17-23, 2008: Proceedings*. Berlin: Springer: 399-410.

Stede, M. 2011. *Discourse Processing*. Synthesis lectures on human language technologies. San Rafael: Morgan & Claypool Publishers.

Tam, Y.-C., Lane, I. & Schultz, T. 2007. Bilingual LSA-Based Adaptation for Statistical Machine Translation. *Machine Translation* 21 (4): 187-207.

Tiedemann, J. 2010a. Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP)*. Stroudsburg: Association for Computational Linguistics: 8-15. Available online: http://aclweb.org/anthology-new/W/W10/W10-2602.pdf.

Tiedemann, J. 2010b. To Cache or Not to Cache? Experiments with Adaptive Models in Statistical Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 189-194. Available online: http://aclweb.org/anthology-new/W/W10/W10-1728.pdf.

Ture, F., Oard, D.W. & Resnik, P. 2012. Encouraging Consistent Translation Choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg: Association for Computational Linguistics: 417-426. Available online: http://aclweb.org/anthology-new/N/N12/N12-1046.pdf.

Vickrey, D. et al. 2005. Word-Sense Disambiguation for Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics: 771-778. Available online: http://www.aclweb.org/anthology-new/H/H05/H05-1097.pdf.

Voigt, R. & Jurafsky, D. 2012. Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Stroudsburg: Association for Computational Linguistics: 18-25. Available online: http://aclweb.org/anthology-new/W/W12/W12-2503.pdf.

Webber, B., Egg, M. & Kordoni, V. 2011. Discourse Structure and Language Technology. *Natural Language Engineering* 18 (4): 437-490.

Wong, B.T.M. & Kit, C. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics: 1060-1068. Available online: http://aclweb.org/anthology-new/D/D12/D12-1097.pdf.

Xiao, T. et al. 2011. Document-Level Consistency Verification in Machine Translation. In *Proceedings of MT Summit XIII*. 131-138. Available online: http://www.mt-archive.info/MTS-2011-Xiao.pdf.

Zhao, B. & Xing, E.P. 2006. BiTAM: Bilingual Topic AdMixture Models for Word Alignment. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Stroudsburg: Association for Computational Linguistics: 969-976. Available online: http://aclweb.org/anthology-new/P/P06/P06-2124.pdf.

Zhao, B. & Xing, E.P. 2008. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In J. Platt et al. (eds.), *Advances in Neural Information Processing Systems* 20. Cambridge (Mass.): MIT Press: 1689-1696.