Machine Translation with Many Manually Labeled Discourse Connectives

Thomas Meyer
Idiap Research Institute and EPFL
Martigny and Lausanne, Switzerland
thomas.meyer@idiap.ch

Lucie Poláková
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Prague, Czech Republic
polakova@ufal.mff.cuni.cz

Abstract

The paper presents machine translation experiments from English to Czech with a large amount of manually annotated discourse connectives. The gold-standard discourse relation annotation leads to better translation performance in ranges of 4–60% for some ambiguous English connectives and helps to find correct syntactical constructs in Czech for less ambiguous connectives. Automatic scoring confirms the stability of the newly built discourse-aware translation systems. Error analysis and human translation evaluation point to the cases where the annotation was most and where less helpful.

1 Introduction

Recently, research in statistical machine translation (SMT) has renewed interest in the fact that for a variety of linguistic phenomena one needs information from a longer-range context. Current statistical translation models and decoding algorithms operate at the sentence and/or phrase level only, not considering already translated context from previous sentences. This local distance is in many cases too restrictive to correctly model lexical cohesion, referential expressions (noun phrases, pronouns), and discourse markers, all of which relate to the sentence(s) before the one to be translated.

Discourse relations between sentences are often conveyed by explicit discourse connectives (DC), such as *although, because, but, since, while*. DCs play a significant role in coherence and readability of a text. Likewise, if a wrong connective is used in translation, the target text can be fully incomprehensible or not conveying the same meaning as was established by the discourse relations in the source text. In English, about 100 types of such explicit connectives have been annotated in the Penn Discourse TreeBank (PDTB, see Section 4), signaling discourse relations such as temporality or contrast between two spans of text. Depending on the set of relations used, there can be up to 130 such relations and combinations thereof. Discourse relations can also be present implicitly (inferred from the context), without any explicit marker being present. Although annotation for implicit DCs exists as well, we only deal with explicit DCs in this paper. DCs are difficult to translate mainly because a same English connective can signal different discourse relations in different contexts and when the target language has either different connectives according to the source relations signaled or uses different lexical or syntactical constructs in place of the English connective.

In this paper, we present MT experiments from English (EN) to Czech (CZ) with a large amount of *manually* annotated DCs. The corpus, the parallel Prague Czech-English Dependency Treebank (PCEDT) (Section 4), is directly usable for MT experiments: the entire discourse annotation in EN is paralleled with a human CZ translation. This means that we can build and evaluate, against the CZ reference, a translation system, that learns from the EN gold standard discourse relations. These then have no distortion from wrongly labeled connectives as it is given in related work (Section 3) where automatic classifiers have been used to label the connectives with a certain error rate. Furthermore, we can use the sense labels for 100 types of EN connectives, whereas related work only focused on a few highly ambiguous connectives that are especially problematic for translation.

The paper starts by illustrating difficult translations involving connectives (Section 2) and discusses related work in Section 3. The resources and data used are introduced in Section 4. The MT experiments are explained in Section 5 and automatic evaluation is given in Section 6. We further provide a detailed manual evaluation and error analysis for the CZ translations generated by our SMT systems (Section 7). Future work described in Section 8 concludes the paper.

2 Motivation

The following example shows a CZ translation of the English DC *meanwhile*. The previous sentences to the example were about other computer producers expected to report disappointing financial results. The interpretation of *meanwhile* and the discourse relation (or sense) signaled is therefore CONTRASTIVE and **not** TEMPORAL:

> **SOURCE**: Apple Computer Inc., **meanwhile**<COMPARISONCONTRAST>, is expected to show improved earnings for the period ended September.
> **BASELINE**: Společnost Apple Computer Inc., **mezitím** by měla ukázat lepší příjmy za období končící v září.
> **SYSTEM2**: Společnost Apple Computer Inc., **naopak** by měla ukázat lepší příjmy za období končící v září.

A baseline SMT system for EN/CZ generated the incorrect CZ connective *mezitím* which signals a temporal relation only. The translation marked SYSTEM2 in the example was output by one of the systems we trained on manual DC annotations (cf. Section 5). The system correctly generated the CZ connective *naopak* signaling a contrastive sense. The example sentence is taken from the Wall Street Journal corpus, section 2365. The sense tag for *meanwhile* was manually annotated in the Penn Discourse TreeBank, see Section 4.

3 Related Work

The disambiguation of DCs can be seen as a special form of Word Sense Disambiguation (WSD), that has been applied to SMT for content words with slight improvements to translation quality (Chan et al., 2007; Carpuat and Wu, 2007). DCs however form a class of procedural function words that relate text spans from an arbitrarily long context and their disambiguation needs features from that longer-range context. Only few studies address function word disambiguation for SMT: Chang et al. (2009) disambiguate a multifunctional Chinese particle for Chinese/English translation and Ma et al. (2011) use tagging of English collocational particles for translation into Chinese. Lexical cohesion at the document level has recently also come into play, with studies on lexical consistency in SMT (Carpuat, 2009; Carpuat and Simard, 2012), topic modeling applied to SMT (Eidelman et al., 2012) or decoding with document-wide features (Hardmeier et al., 2012). A recently published article summarizes most of the work on SMT with the broader perspective of discourse, lexical cohesion and coreference (Hardmeier, 2013).

For discourse relations and DCs especially, more and more annotated resources have become available in several languages, such as English (Prasad et al., 2008), French (Péry-Woodley et al., 2009; Danlos et al., 2012), German (Stede, 2004), Arabic (AlSaif, 2012), Chinese (Zhou and Xue, 2012) and Czech (Mladová et al., 2009). These resources however remain mostly monolingual, i.e. translations or parallel texts in other languages do normally not exist. This makes these resources not directly usable for MT experiments.

Recent work has shown that more adequate and coherent translations can be generated for English/French when ambiguous connectives in the source language are annotated with the discourse relation they signal (Popescu-Belis et al., 2012). SMT systems for European language pairs are most often trained on Europarl corpus data (Koehn, 2005), where only a small amount of discourse-annotated instances is available (8 connectives with about 300-500 manual annotations each). Meyer and Popescu-Belis (2012) therefore used these few examples to train automatic classifiers that introduce the sense labels for the connectives in the entire English text of the Europarl corpus. Although these classifiers are state-of-the-art, they can have an error rate of up to 30% when labeling unseen instances of connectives. The discourse-aware SMT systems nevertheless improved about 8-10% of the connective translations. When integrating into SMT directly the small manually-labeled data, without training classifiers, hardly any translation improvement was measurable, cf. (Meyer and Popescu-Belis, 2012).

4 The Parallel Prague Czech-English Dependency Treebank

With the English-Czech parallel text provided in the Prague Czech-English Dependency Treebank 2.0 (PCEDT) (Hajič et al., 2011)[1], comes a human CZ translation of the entire Wall Street Journal Corpus in EN (WSJ, sections 00-24, approxi-

[1] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T08

mately 50k sentences).

The syntactical annotation of WSJ, the Penn TreeBank (Marcus et al., 1993), has been followed by a discourse annotation project, the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), over the same sections of the corpus. In the PDTB version 2.0, 18,459 instances of explicit DCs, among other discourse-related phenomena (implicit relations, alternative lexicalizations), are labeled along with the text spans they connect (discourse arguments) and the discourse relation they signal (sense tags).

The sense tags are organized in a three-level sense hierarchy with four top semantic classes, 16 sub-senses on the second and further 23 subsenses on the third hierarchy level. The annotators were not forced to make the finest distinction (on the sub-sense level). A token can also be annotated with two senses, forming a composite sense with a label combination from wherever in the hierarchy, resulting in 129 theoretically possible distinct sense tags (see Section 5 for the sense levels we use). For the latter reason, some of the sense labels are very scarcely used and although they make for important and fine-grained distinctions in English, this granularity level might not be useful for translation, where only certain ambiguities have to be resolved to obtain a correct target language connective, see Section 7.

The PCEDT is a 1:1 sentence-aligned parallel resource with a manual multilayer dependency analysis of both original Penn TreeBank-WSJ texts and their translations to Czech. Despite the manually annotated parallel dependency trees which are very valuable in other linguistic studies, for translation we only used the plain CZ texts provided with the treebank.

5 Experimental Setup

In the following, we describe a series of SMT experiments that made direct use of the EN/CZ text as provided with the PCEDT. The SMT models were all phrase-based and trained with the Moses decoder (Koehn et al., 2007), either on plain text for the BASELINE or on text where the EN connective word-forms have been concatenated with the PDTB sense labels. All texts have been tokenized and lowercased with the Moses tools before training SMT. In future work, we will build factored translation models (Koehn and Hoang, 2007) as well, as this would reduce the label scarcity that was likely a problem when just concatenating word-forms and labels (see Sections 7 and 8).

For SYSTEM1 in the following, we inserted, into the English side of the PCEDT data, the full sense labels from the PDTB, which can be, as already mentioned, as detailed as containing 3 sense levels and allowing for composite tags (where annotators chose that two senses hold at the same time). SYSTEM1 therefore operates on a total of 63 distinct and observed sense tags for all DCs.

For SYSTEM2, we reduced the sense labels to contain only senses from PDTB sense hierarchy level 2 and 1, not allowing for composite senses, i.e. for those instances that were annotated with two senses we discarded the secondary (but not less important) sense. This reduced the set of senses for SYSTEM2 to 22.

The procedure is exemplified in the example below with an EN sentence 1 (WSJ section 2300) containing a complex PDTB sense tag that has been kept for SYSTEM1. For SYSTEM2 we have reduced the sense of *when* to: <ContingencyConditionGeneral>. Sentence 2 (WSJ section 2341) contains two already simplified sense tags. The original PDTB sense tags for *meanwhile* and *as* were respectively <ComparisonContrastJuxtaposition> and <ContingencyPragmaticcause-Justification>, where Juxtaposition and Justification were dropped because they stem from the third level of the PDTB sense hierarchy:

1. Selling snowballed because of waves of automatic "stop-loss" orders, which are triggered by computer **when**<ContingencyConditionGeneral-TemporalAsynchronousSuccession> prices fall to certain levels.
2. Meanwhile<ComparisonContrast>, analysts said Pfizer's recent string of lackluster quarterly performances continued, **as**<ContingencyPragmaticcause> earnings in the quarter were expected to decline by about 5%.

In order to build SMT systems of reasonable quality, we still need to combine the PCEDT texts (50k sentences) with other resources such as the EN/CZ parts of the Europarl corpus. This results in a mixture of labeled and unlabeled DCs in the data and estimates might be noisy. We however also checked system performance on the PDTB test set (section 23) with labeled DCs only (see Section 6) for which the unlabeled ones in the model do not pose a problem, as they are not considered as valid target phrases by the SMT decoder. The following list gives an overview of the data used to build three SMT systems. No modifications have been done to the texts of the BASELINE system, that uses exactly the same amount of sentences, but no sense labels.

- BASELINE: no tags for connectives
- SYSTEM1: complex PDTB sense tags
- SYSTEM2: simplified PDTB sense tags
- training: Europarlv7 (645,155 sentences) + PDTB sections 02-21 (41,532 sentences; 15,402 connectives)
- tuning: newstest2011 (3,003 sentences) + PDTB sections 00,01,22,24 (5,260 sentences; 2,134 connectives)
- testing: newstest2012 (3,001 sentences) + PDTB section 23 (2,416 sentences; 923 connectives)[2]

The language model, the same for BASELINE, SYSTEM1 and SYSTEM2, was built using SRILM (Stolcke et al., 2011) with 5-grams over Europarl and the news data sets 2007-2011 in CZ, as distributed by the Workshop on Machine Translation[3]. All systems were tuned by MERT (Och, 2003) as implemented in Moses.

6 Automatic Evaluation

Most automatic MT scoring relies on n-gram matching of a system's candidate translation against (usually) only one human reference translation. For DCs therefore, automatic scores do not reveal much of a system's performance, as often only one or two words, i.e. the DC is changed. When a candidate translation however contains a more accurate and correct connective, the translation output is often more coherent and readable than the baseline's output, see Section 7.

Automatic evaluation has been done using the MultEval tool, version 0.5.1 (Clark et al., 2011). The BLEU scores are computed by jBLEU V0.1.1 (an exact reimplementation of NIST's mteval-v13.pl without tokenization). Table 1 provides an overview of the BLEU scores for the BASELINE and systems 1 and 2 on the full test set (newstest2012 + PDTB section 23), and on PDTB section 23 only, the latter containing 2,416 sentences and 923 labeled DCs.

In order to gain reliable automatic evaluation scores, we executed 5 runs of MERT for each translation model configuration. MERT is implemented as a randomized, non-deterministic optimization process, so that each run leads to different feature weights and as a consequence, to different BLEU scores when translating unseen text. The scores from the 5 runs were then averaged and with a t-test we calculated the confidence p-values for the score differences. When these are below 0.05, they confirm that it is statistically likely, that such scores would occur again in other tuning runs. In terms of BLEU, neither SYSTEM1 nor SYSTEM2 therefore performs significantly better or worse than the BASELINE.

In order to show how little the DC labeling actually affects the BLEU score, we randomized all connective sense tags in PDTB test section 23 and translated again 5 times (with the weights from each tuning run) with both, SYSTEM1 and SYSTEM2. With randomized labels, both systems perform statistically significantly worse ($p = 0.01$, marked with a star in Table 1) than the BASELINE, but only with an average performance loss of -0.6 BLEU points. Note that some sense tags might still have been correct due to randomization.

Test set	System	BLEU
nt2012 + PDTB 23	BASELINE	17.6
	SYSTEM1	17.6
	SYSTEM2	17.6
PDTB 23	BASELINE	21.4
	SYSTEM1	21.4
	SYSTEM2	21.4
PDTB 23 random	SYSTEM1	20.8*
	SYSTEM2	20.8*

Table 1: BLEU scores when testing on the combined test set (newstest2012 + PDTB 23); on PDTB section 23 only (2416 sentences, 923 connectives); and when randomizing the sense tags (PDTB 23 random), for the BASELINE system and the two systems using PDTB connective labels: SYSTEM1: complex labels, SYSTEM2: simplified labels. When testing on randomized sense labels (PDTB 23 random), the BLEU scores are statistically significantly lower than the ones on the correctly labeled test set (PDTB 23), which is indicated by starred values.

Automatic MT scoring does therefore not reveal actual changes in translation quality due to DC usage. In the next section, we manually analyze

[2] Note that this PDTB section division for training, development and testing is the same as is used for automatic classification experiments, as recommended in the PDTB annotation manual.

[3] http://www.statmt.org/wmt12/

samples of the translation output by SYSTEM2 that reached the highest scores observed in some of the single tuning runs before averaging.

7 Manual Evaluation and Error Analysis

Two human judges went both through two random samples of SYSTEM2 translations from WSJ section 23, namely sentences 1-300 and 1000-2416. In these sentences, there were 630 observed connectives. The judges counted the translations that were better, equal and worse in terms of the DCs as output by SYSTEM2 versus the BASELINE system. We then summarized the counts over the two samples and give the scores as $\Delta(\%)$ in Table 2. To further test if we just had bad samples, the judges went through another set of translations (1024–1138), containing 50 DCs, for which the counts are summarized in Table 2 as well. A translation was counted as being correct when it generated a valid CZ connective for the corresponding context, without grading the rest of the sentences.

Overall, it was found that the number of better translations is only slightly higher for SYSTEM2 than the ones from the BASELINE system. The vast majority of DCs was translated correctly by both the BASELINE and SYSTEM2, and in very few cases, both systems translated the DCs incorrectly.

SYSTEM2 appeared to systematically repeat one mistake, namely translating the very frequent connective *but* preferably with *jenže*, which is correct but rare in CZ (the primary and default equivalent for *but* in CZ is *ale*). This 'mis-learning' likely happened to a frequent correspondence of *but–jenže* in the SMT training data, which then does not necessarily scale to and be of appropriate style in the testing data. If one disregards these occurrences, SYSTEM2 translates between about 8 and 20% of all connectives better than the BASELINE (discounted percentages for *jenže* in Table 2). The results seem therefore to be dependent on the parts of the test set evaluated and the DCs occurring in them.

The only slight quantitative improvements and cases were SYSTEM2 performed worse are most likely due to the overall scarcity of the PDTB sense tags (cf. Section 4). Especially for SYSTEM1 but to some extent also for SYSTEM2, rare sense tags such as CONTINGENCYPRAGMATICCAUSE might not be seen often or even not at all in the SMT training data and therefore not be learned appropriately to provide good translations for the test data. In relation to that, simply concatenating the sense tags onto the connective word-forms leads to scarcity of the latter, whereas other ways to include linguistic labels in SMT, such as factored translation models, would account for the labels as additional translation features, which will be investigated in future work (Section 8).

In the following, we analyze cases where SYSTEM2 translates the connectives better and more appropriately than the BASELINE. These cases include highly ambiguous connectives, temporal DCs with verbal *ing*-forms and conditionals.

In general, for the very ambiguous EN connectives (e.g. *as, when, while)*, disambiguated for SYSTEM2 with the PDTB sense tags, we indeed obtained more accurate translations than those generated by the BASELINE. One of the human judges had a close look at 25 randomly sampled instances of *as*, taken from the manually evaluated sets mentioned above. In these test cases, 68% of all occurrences of *as* were better translated by SYSTEM2 and only 4% of the translations were degraded when compared to the BASELINE. For details, see Table 3[4].

In the following translation example (WSJ section 2365), and often elsewhere, the BASELINE system treats the connective *as* as a preposition *jako* with the meaning *She worked as a teacher.* This frequent interpretation seems to be learned quite reasonably from the SMT training data, it is however incorrect where *as* actually functions as a DC. SYSTEM2, in agreement with the tagging, then correctly generates the causal connective *protože*:

SOURCE: In the occupied lands, underground leaders of the Arab uprising rejected a U.S. plan to arrange Israeli-Palestinian talks **as**<CONTINGENCYCAUSE> Shamir opposed holding such discussions in Cairo.
BASELINE: *Na okupovaných územích, podzemní vůdců arabských povstání odmítl americký plán uspořádat izraelsko-palestinské rozhovory **jako** Šamira proti pořádání takových diskusí v Káhiře.
SYSTEM2: Na okupovaných územích, podzemní vůdců arabského povstání odmítl americký plán uspořádat izraelsko-palestinské rozhovory, **protože** Šamira proti pořádání takových diskusí v Káhiře.

DCs can also be translated to other syntactical constructs available in the target language that convey the same discourse relation without any

[4] We included simple occurrences only, i.e. not compound connectives like *as if, as soon as* or translations were the connective was dropped. In the PDTB, *as* can have up to 17 distinct senses, ranging from temporal, causal to concessive relations.

Configuration	Δ(%) **vs. BASELINE**			**Total (%)**
	Improved	Equal	Degraded	
sentences 1–300 / 1000–2416 630 labeled DCs				
SYSTEM2	7.9	75.2	9.4	92.5
not counting 25 x *but–jenže*	8.2	80.3	4.0	92.5
both systems wrong				7.5
				100
sentences 1024–1138 50 labeled DCs				
SYSTEM2	16	76	6	98
not counting 2 x *but–jenže*	19	77	2	98
both systems wrong				2
				100

Table 2: Performance of SYSTEM2 (simplified PDTB tags) when manually counting for improved, equal and degraded translations compared to the BASELINE, in samples from the PDTB section 23 test set.

explicit DC. For EN/CZ this occurs for DCs such as *before/after/since* + Verb in Present Continuous. In CZ, these either should be rendered as a verbal clause or a nominalization. We accounted for translations as being well-formed, if the SMT systems generated one of these possibilities correctly, i.e. not only the connective/preposition but also the verb/noun. In CZ, it must be decided between using a preposition (e.g. *před*) or a connective (e.g. *než*). A good translation would for example be: *before climbing* = PREP+NP or DC+V, and a bad translation: *before climbing* = PREP+V/ADJ or DC+NP. The following example (WSJ section 2381) is a SYSTEM2 output where the sense tag in English helped to translate the connective *before* more correctly by DC+V, whereas the BASELINE renders this wrongly by using PREP+ADJ:

SOURCE: Mr. Weisman predicts stocks will appear to stabilize in the next few days **before**<TEMPORALASYNCHRONOUS> declining again, trapping more investors.
BASELINE: *Pan Weisman předpovídá, že akcie budou stabilizovat v příštích několika dnech před/**PREP** klesajícím/**ADJ** opět odchytu více investorů.
SYSTEM2: Pan Weisman předpovídá, že akcie bude stabilizovat, jak se zdá, v příštích několika dní, než/**DC** opět klesat/**V**, zablokování více investorů.

A further difficult case in CZ is the binding of conditionals with personal pronouns, e.g. *if I* = *kdybych*, *if you* = *kdybys*, *if he/she* = *kdyby* etc. In the following example (WSJ section 2386), the BASELINE system completely missed to render the personal pronoun (but still generated the correct conditional connective *if–pokud*), whereas SYSTEM2 outputs the much better *if I–kdybych*. However, apart from the better connective, SYSTEM2's translation is worse than the BASELINE's, because the first verb form is misconjugated and the second verb (*will take*) is missing:

SOURCE: **If**<CONTINGENCYCONDITION> I sell now, I'll take a big loss.
BASELINE: ***Pokud** chtěl prodat, teď budu brát s velkou ztrátou.
LIT.: If he-wanted to-sell, now I-will take with big-Instrumental loss-Instrumental.
SYSTEM2: **Kdybych** se nyní prodávají, se z tohohle velkou ztrátu.
LIT.: If-I themselves-ReflexPron now they-are selling, ReflexPron out-of this big-Accusative loss-Accusative.

From the automatic and manual translation evaluation, we conclude that using the sense tags for *all* 100 connectives in EN is not the most appropriate method, and that only certain connectives such as *as, when, while, yet* and a few others are very problematic in translation due to the many discourse relations they can signal. In future work, we will therefore analyze in more detail which connectives and which sense labels from the PDTB should actually be included in the data to train SMT.

BASELINE	SYSTEM2	occ.	PDTB
jak	když	1	Sy
jak	**když**	1	Sy
jelikož	jelikož	1	Ca
neboť	neboť	1	Ca
protože	protože	2	Sy/Co; Ca
a	protože	1	Sy/Co
aby	když	1	Sy
jak	když	1	Sy
jak	protože	1	Ca
jako	protože	4	Sy/Co; Ca
jako	když	5	Sy; Asy; Ca
jako	kdy	2	Sy
protože	když	1	Sy
že	když	1	Sy
jako	**jak**	1	Sy
jako	**poté, co**	1	Sy
Total		25	
sys2 +		68%	
sys2 =		20%	
sys2 –		4%	
both –		8%	

Table 3: Translation outputs for the EN connective *as*, which was translated more correctly by SYSTEM2 thanks to the disambiguating sense tags compared to the BASELINE that often just produces the prepositional *as – jako*. The erroneous translations are marked in bold. The PDTB sense tags indicate the meaning of the CZ translations and are encoded as follows: Synchrony (Sy), Asynchrony (Asy), Contingency (Co), Cause (Ca).

8 Conclusion

We presented experiments for EN/CZ SMT with a large amount of hand-labeled discourse connectives that are disambiguated in the source language and training material for MT systems by their sense tags or discourse relations they signal. This leads to improved translations in cases where the source DC is highly ambiguous or where the target language uses other syntactical constructs than a connective to convey the discourse relation.

Using all 100 types of EN DCs in the corpus and/or all the detailed sense tags from the manual annotation most probably lead to the only very slight improvements for the discourse-aware systems when measured quantitatively over the whole test sets. In future work we plan to more thoroughly analyze which connectives need to be disambiguated at which sense granularity level before implementing them into an SMT system.

For label implementation there also are other ways worth examining, such as factored translation models that handle the supplementary linguistic information as separate features and alternative decoding paths.

Acknowledgments

We are grateful for the funding of this work to the Swiss National Science Foundation (SNSF) under the COMTIS Sinergia project, n. CRSI22_127510 (see `www.idiap.ch/comtis/`), to the Grant Agency of the Czech Republic (project n. P406/12/0658) and to the SVV of the Charles University (project n. 267 314). We would like to thank Lenka Sándor for her help with manual translation evaluation.

References

Amal AlSaif. 2012. *Human and Automatic Annotation of Discourse Relations for Arabic*. Ph.D. thesis, University of Leeds.

Marine Carpuat and Michel Simard. 2012. The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, pages 442–449, Montreal, Canada.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 61–72, Prague, Czech Republic.

Marine Carpuat. 2009. One Translation per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW)*, pages 19–27, Singapore.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 33–40, Prague, Czech Republic.

Pi-Chuan Chang, Dan Jurafsky, and Christopher D. Manning. 2009. Disambiguating 'DE' for Chinese-English Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, Athens, Greece.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies)*, Portland, OR.

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. 2012. Vers le FDTB : French Discourse Tree Bank. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 471–478, Grenoble, France.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of ACL 2012 (50th Annual Meeting of the Association for Computational Linguistics*, pages 115–119, Jeju, Republic of Korea.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2011. Prague Czech-English Dependency Treebank 2.0. Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju, Korea.

Christian Hardmeier. 2013. Discourse in Statistical Machine Translation. *DISCOURS*, 11:1–29.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, pages 868–876, Prague, Czech Republic.

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 45th Annual Meeting of the Association for Computational Linguistics (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*, pages 79–86, Phuket, Thailand.

Jianjun Ma, Degen Huang, Haixia Liu, and Wenfeng Sheng. 2011. POS Tagging of English Particles for Machine Translation. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 57–63, Xiamen, China.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 129–138, Avignon, FR.

Lucie Mladová, Šárka Zikánová, Zuzanna Bedřichová, and Eva Hajičová. 2009. Towards a discourse corpus of Czech. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Marie-Paule Péry-Woodley, Nicholas Asher, Patrice Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, Philippe Muller, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, Laure Vieu, and Antoine Widlöcher. 2009. ANNODIS: une approche outillée de l'annotation de structures discursives. In *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Paris, France.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

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 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech, Morocco.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (ACL)*, Jeju Island, Korea.