Comparison of Coreference Resolvers for Deep Syntax Translation [*]

Michal Novák,[*] **Dieke Oele,**[‡] **Gertjan van Noord,**[‡]
[*]Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
`mnovak@ufal.mff.cuni.cz`
[‡]University of Groningen, The Netherlands
`{d.oele,g.j.m.van.noord}@rug.nl`

Abstract

This work focuses on using anaphora for machine translation with deep-syntactic transfer. We compare multiple coreference resolvers for English in terms of how they affect the quality of pronoun translation in English-Czech and English-Dutch machine translation systems with deep transfer. We examine which pronouns in the target language depend on anaphoric information, and design rules that take advantage of this information. The resolvers' performance measured by translation quality is contrasted with their intrinsic evaluation results. In addition, a more detailed manual analysis of English-to-Czech translation was carried out.

1 Introduction

Over the last years, the interest in addressing coreference-related issues in Machine Translation (MT) has increased. Multiple works focused on using information coming from a Coreference Resolution (CR) system to improve pronoun translation in phrase-based frameworks (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012). A similar task was addressed in the TectoMT deep syntax tree-to-tree translation system (Žabokrtský et al., 2008). Novák et al. (2013a; 2013b) presented specialized models for the personal pronoun *it* and reflexive pronouns in English-Czech translation, which resulted in an improvement in terms of human evaluation. Although these models were tailored to pronoun translation, they only addressed cases

where anaphora information is in fact not needed. For proper translation of other pronouns, however, coreference must be involved.

The present work concentrates on exploiting coreference for deep syntax MT. We integrate three coreference resolvers for English into TectoMT system, namely the Treex CR (Popel and Žabokrtský, 2010), the Stanford CR (Lee et al., 2013), and BART (Versley et al., 2008), and observe their effects on translation quality. Taking linguistic observations on the target languages into account, we design rules that make use of the information supplied by the CR systems. We apply this approach to English-Czech and English-Dutch translation.

This paper is structured as follows. In Section 2, we introduce the grammar of Czech and Dutch pronouns with a special emphasis on cases where form depends on anaphoric relations. Section 3 gives a brief description of the used CR systems. In Section 4, the TectoMT system is presented, along with our rules exploiting coreference information. In Section 5, the individual configurations of the MT system are evaluated using BLEU score and human evaluation. These evaluations are contrasted with intrinsic scores of the CR systems. The results of English-to-Czech translation are analyzed in a greater detail in Section 6. Ultimately, this paper is concluded in Section 7.

2 Pronouns in the target languages

The system of anaphoric pronouns is similar for Czech and Dutch, both containing personal, possessive, reflexive, relative, and demonstrative pronouns.[1]

In the present work, we mainly concentrate on a subset of anaphoric pronouns whose form cannot be reliably determined without knowing the closest co-referring mention (the antecedent) and its

[1]We omit demonstrative pronouns in this work as they are not consistently treated by any of the CR systems used here.

grammatical properties. Grammatical properties (such as morphological gender, number, or syntactic position) and the agreement of such pronoun and its antecedent are the key factors that suggest to the reader which entity the pronoun refers to.

2.1 Pronouns in Czech

The typology of Czech pronouns is the following:

Personal pronouns. Their form depends on the person (cf. *já /I/* in 1st person and *ty /you/* in 2nd person), number (cf. *já /I/* in singular and *my /we/* in plural), gender (cf. masculine *on /he/*, feminine *ona /she/* and neuter *ono /it/*), and case (cf. *on /he/* in nominative and *jemu /to him/* in dative). In addition, some forms may have a short or a long variant. As Czech is a pro-drop language, pronouns in the subject can be even omitted from the surface.[2] Out of these features, only gender and number depend on anaphora – they must agree with antecedent's gender and number.

Possessive pronouns. Unlike personal pronouns, two types of gender and number are distinguished for possessives: one agreeing with the possessed object and one agreeing with the possessor (cf. feminine–masculine *s jeho ženou /with his wife/*, masculine–feminine *s jejím mužem /with her husband/* and feminine–feminine *s její ženou /with her wife/*). The latter type of gender and number depends on anaphora, as the antecedent of the possessive pronoun is in fact their possessor.

Reflexive pronouns. Their form is determined only by the case and variant. Unlike English, they carry no gender and number information. However, information on anaphora is still required to specify whether a reflexive or a personal pronoun should be used, since reflexive pronoun are used in case of coreference with the sentence subject.

Reflexive possessive pronouns. A special category of pronouns which is used instead of a possessive pronoun if its possessor is the sentence subject. They do not depend on other grammatical features of the antecedent than its syntactic position, as reflexives do.

Relative pronouns. Relative pronouns need to agree in gender and number[3] with their antecedent. However, their usage is limited by the

nature of their antecedents, e.g., while the pronoun *který /which, that, who/* can be used in most cases where the antecedent is a noun phrase, the pronoun *což /which/* is required whenever referring to a clause or a longer utterance.

2.2 Pronouns in Dutch

The typology of Dutch anaphoric pronouns is the following:

Personal pronouns. The form of Dutch personal pronouns depends on person, case, number, and gender. They are used in a similar way as in English. Nouns are partitioned by the article they use: *de* or *het*. While *het-* nouns are referred to by the pronoun *het*, masculine pronouns are mostly used for *de-* nouns. Feminine pronouns can be used for abstract feminine nouns.

Possessive pronouns. Possessive pronouns differ with respect to person, gender, and number. In addition, some of them agree in gender with their head noun (e.g., *ons* versus *onze /our/*). They make no distinction based on whether they refer to a *het-* or *de-* noun.

Reflexive pronouns. Each Dutch personal pronoun has a reflexive form that can differ with respect to person and number, but not gender, i.e. *zich/zichzelf* is used for all genders.

Relative pronouns. The distinction between relative pronouns *die* and *dat* is determined by the type of the antecedent. *Die* is used when it refers to a *de-* noun or any plural form while *dat* is used for singular *het-* nouns. When the pronoun refers to a person with a direct object function, *wie* is used. *Wat* can refer to indefinite words, superlatives, or whole phrases while *welke* can solely refer to *de-* words but is mostly used in formal texts. A relative pronoun turns into a so-called pronominal adverb if it is part of a prepositional phrase (e.g., preposition+*die/dat* is replaced by *waar*+preposition).

3 Coreference resolution systems for English

We apply the Treex CR system, the BART system and the Stanford Deterministic CR system in our experiments[4]. As neither BART nor the Stanford

---

[2]In that case, some of the pronoun's properties can be reconstructed from the verb thanks to subject-verb agreement.

[3]Possessor's gender and number in case of possessive relative pronouns.

[4]The reasons for choosing the latter two systems are twofold: they are freely available and they perform close to the state of the art, as confirmed by the results of CoNLL-2012 Shared Task (Pradhan et al., 2012).

system target relative pronouns, we combine these two systems with a Treex module for relative pronouns (the *Treex-relat* module).

3.1 Treex Coreference Resolution System

This system is a part of the Treex framework (Popel and Žabokrtský, 2010) and has been used for English-to-Czech translation in the TectoMT system (Žabokrtský et al., 2008). It consists of several modules; each of them focuses on a specific type of coreferential relations in English:[5] anaphora of relative pronouns (the *Treex-relat* module) and personal, possessive, and reflexive pronouns (the *Treex-other* module). All the modules are rule-based, making use of syntactic representation of the sentence as well as simple context heuristics.

3.2 BART

BART 2.0 (Versley et al., 2008; Uryupina et al., 2012) is a modular toolkit for end-to-end coreference resolution. It is based on mention-pair model, which means that a classifier makes a decision for every pair of mentions whether they belong to the same coreference cluster or not. Subsequently, the mentions paired by pairwise decisions need to be partitioned into coreference chains. The model is trained using the WEKA machine-learning toolkit (Witten and Frank, 2005). Features for English are identical to those used in virtually all state-of-the-art coreference resolvers (Soon et al., 2001).

3.3 Stanford Deterministic Coreference Resolution System

The Stanford resolver (Lee et al., 2013) is a state-of-the-art rule-based system. Unlike BART, it is an entity-based system, meaning that in each step, the system decides on assigning a mention into one of the partially created coreference chains. It proceeds in multiple steps – *sieves*, starting with high-precision rules and ending with those with a lower precision but a higher recall. The version of the system used here consist of ten sieves including the sieve for pronominal mentions in quotations, sieves for string match, head match, proper head noun match, and the pronoun match applied at the end.

---

[5]A similar system for Czech pronouns is also a part of the Treex framework.

4 The TectoMT System and Coreference

TectoMT (Žabokrtský et al., 2008) is a tree-to-tree machine translation system whose translation process follows the analysis-transfer-synthesis pipeline.

In the analysis stage, the source sentence is transformed into a deep syntax dependency representation based on the Prague tectogrammatics theory (Sgall et al., 1986). At this point, CR systems are applied to interlink the tree representation with coreference relations.

The source language tree structure is transferred to the target language using three factors: translation models for deep lemmas, morpho-syntactic form labels, and a rule-based factor for other grammatical properties. For the most part, isomorphism of the tree representation in both languages is assumed, and the tree is translated node-by-node.

In the last step, the deep representation is transformed to a surface sentence in the target language.

English-to-Czech translation has been developed and tuned in TectoMT since its very beginning. Translation to other languages, including Dutch, was added only recently (Popel et al., 2015).

4.1 Rules Using Coreference

During the transfer and the synthesis stage, language-dependent rules that make use of the projected coreference relations are applied. The rules are based on linguistic observations presented in Section 2.

Even if a given grammatical property is ruled by the antecedent, it is not always necessary to use anaphora information. The correct form in the target language can be inferred from the source language word itself. For example, genders in English and Czech are of a different nature. While the gender of English pronouns is *notional*, reserving masculine and feminine gender exclusively for persons (Quirk et al., 1985), the Czech gender is *grammatical* with all gender values more evenly distributed. However, masculine and feminine pronouns mostly remain the same in English-to-Czech translation. Other similar phenomena can be observed in both Czech and Dutch.

Czech. Rules employing coreference resolution have been used in TectoMT English-Czech translation since its beginning, but their contribution

has not been evaluated so far. The following rules are used:

- Impose agreement in gender and number for personal, possessive, and relative pronouns translated from English pronouns *it* and *its* as well as English relative pronouns.[6]
- Transform a possessive to a reflexive possessive pronoun if it refers to a sentence subject.
- Transform a relative pronoun referring to a verb phrase into the Czech relative pronoun *což*.

Dutch. In translation to Dutch, possessives can be inferred solely using the source pronoun. Therefore, only personal and relative pronouns are targeted with the following coreference-based rules:

- Impose agreement in gender (*het-* or *de-* type) for personal pronouns translated from the English pronoun *it*.
- For relative pronouns, a corresponding form is picked based on whether the pronoun is bound in a prepositional phrase, refers to a verb phrase, a person, or a *het-* or *de-* noun.

5   Automatic evaluation

The TectoMT translation models were trained on parallel data from CzEng 1.0 (Bojar et al., 2012) and a concatenation of Europarl (Koehn, 2005), Dutch parallel corpus (Macken et al., 2007) and KDE4 localizations (Tiedemann, 2009), for Czech and Dutch, respectively.

We tested the English-Czech and English-Dutch translation systems on datasets from two different domains: the news domain, represented by English-Czech test set for the WMT 2012 Translation Task (Callison-Burch et al., 2012) as well as the last 36 documents from English-Dutch News Commentary data set (Tiedemann, 2012),[7] and the IT domain, represented by the corresponding pairs of the QTLeap Corpus Batch 2 (Osenova et al., 2015).[8]

The evaluation was conducted for several configurations of TectoMT. The *Baseline* systems did not use any coreference-related rules while the remaining configurations apply all TectoMT coref-

---

[6]For other English pronouns, it appeared to be sufficient to copy their gender and number to the translated pronoun.

[7]Since the original dataset contains a substantial amount of neighboring words stuck together, we corrected it using a spellchecker.

[8]http://metashare.metanet4u.eu/go2/qtleapcorpus

erence rules. They combine the Treex CR module for relative pronouns *Treex-relat* with the three resolvers detailed in Section 3: the *Treex-other* module, the Stanford system, and the BART system. Table 1 shows BLEU scores of all four configurations with respect to the domain and the target language. In addition, it presents an intrinsic evaluation of the CR – anaphora resolution F-scores measured on English parts of sections 20–21 of the Prague Czech-English Dependency Treebank (Hajič et al., 2012).

The results reveal that for every domain and language, there is at least a single coreference-aware configuration that outperforms the baseline.

In addition to the substantial BLEU difference between the Czech best system and the baseline, all the coreference-aware configurations improved upon the baseline translation into Czech. On the other hand, we observed a very small improvement of the best Dutch system over the baseline.

This disproportion reflects the fact that whereas English-to-Czech TectoMT has been developed and tuned over seven years, the English-to-Dutch translation was added only recently.

Comparable scores of Czech systems on the IT domain can be attributed to two aspects: TectoMT has mostly been tuned to the news domain, and the distribution of pronoun types may differ.

When contrasting BLEU scores with the intrinsic evalution of CR systems, one can see that although their performance is similar, their effect on translation quality varies across languages and domains. The results also show that out of all pronoun types, CR of relative pronouns is the most reliable. This is confirmed by consistent gains of the *Treex-relat* system over the baseline.

6   Manual analysis of the results

BLEU score has previously been shown not to be suitable for measuring small modifications such as changes in pronouns (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012; Hardmeier, 2014). Despite these findings, we succeeded in getting a better BLEU score with coreference-aware systems for English-to-Czech translation. To reveal a reason for such behaviour, we conducted a detailed analysis of the translation results on the English-Czech news domain dataset.

The data-set comprising almost 64,000 English words contains 894 occurrences are relative pronouns, 770 possessive pronouns and 1950

	Czech		Dutch		Intrinsic			
	news	IT	news	IT	pers	poss	relat	total
Baseline	11.12	30.55	11.76	24.22		—		
Treex-relat	11.45	31.08	**11.78**	**24.25**	–	–	73.64	
Treex-relat+other	**11.54**	31.08	10.55	24.06	54.05	64.09	73.64	62.78
Stanford + Treex-relat	11.45	**31.10**	11.79	24.22	54.08	57.20	73.64	60.65
BART + Treex-relat	11.48	31.09	11.76	24.17	56.61	60.02	73.64	62.45

Table 1: BLEU scores of the TecotMT system for English-Czech and English-Dutch translation using various CR systems, contrasted with their intrinsic evaluation measured by F-score.

	pers	poss	relat
Treex-relat	–	–	337
Treex-relat+other	0	339	337
Stanford + Treex-relat	40	44	337
BART + Treex-relat	128	188	337
potential	1950	770	894

Table 2: Number of changed Czech pronoun translations if the Baseline system is replaced by each of the coreference-aware systems. The last line indicates number of English pronouns of a particular type in the dataset.

personal pronouns. Table 2 presents in how many cases the translation of these pronouns was changed when the baseline system was replaced by each of the coreference-aware systems. [9]

Not surprisingly, all the systems produced exactly the same amount of changes in relative pronouns. We randomly sampled and manually inspected 30 translation changes.[10] Most of the changes are caused by imposing agreement between the pronoun and its antecedent. Compared with the Baseline, in 24 cases the output is better, it is worse in 3 cases and in 3 cases equally bad. In 12 of the improved cases, the produced form of the Czech relative pronoun matches a unigram in the reference translation. Since the relative pronouns are often subjects of the clause, the form of the governing verb is also affected due to agreement rules in Czech. This typically results in matches longer than unigrams, justifying the BLEU score improvement.

Regarding possessive pronouns, the Stanford coreference resolver seems to be very conservative explaining the lack of change in BLEU score

compared to the Treex-relat system. We sampled 30 changes of the Treex-relat+other system and observed 16 better, 7 worse, 4 equally good and 3 equally bad translations compared to the Baseline system. Most of the changes stem from the transformation of possessives to reflexive possessives. The improvement is less convincing for relative pronouns, which correlates with the measured BLEU scores.

As for personal pronouns, the Stanford system confirmed its conservative nature while surprisingly, the Treex coreference system produced no change at all. We sampled 30 translations produced by BART+Treex-relat system and compared it with the Baseline system: 14 translation were better, 7 worse, and 9 equally bad.

For English-to-Dutch translation, we carried out human evaluation with no pronoun type distinction, comparing 30 changed sentences randomly selected from the news domain dataset. The best system's output was considered better in 13 cases, worse in 11 cases, confirming the marginal BLEU changes.

7 Conclusion

In this work, we compared three systems for coreference resolution with regard to what effect they have on the quality of deep syntax machine translation. We found that the results are heavily affected by the quality of the used coreference rules as well as by the language they are applied to. While coreference is essential for better results in the well-tuned translation into Czech, it is so far disputable in translation into Dutch. The reliability of coreference resolution also plays a key role there as the most reliable resolver for relative pronouns was the only one that consistently improved the translation. Manual analysis of the results confirmed the outcomes of the automatic evaluation.

---

[9]If the subject pronoun is dropped from the surface, we decide on the verb properties.

[10]In manual evaluation, the source, automatic and reference translations of the current and two previous sentences were presented to the human judge.

References

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The joy of parallelism with CzEng 1.0. In *Proceedings of LREC 2012*. European Language Resources Association.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51. Association for Computational Linguistics.

Liane Guillou. 2012. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the EACL*, pages 1–10. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289. ISCA.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Comput. Linguist.*, 39(4):885–916.

Lieve Macken, Julia Trushkina, and Lidia Rura. 2007. Dutch parallel corpus: Mt corpus and translator's aid. In *In Proceedings of the Machine Translation Summit XI*, pages 313–320. European Association for Machine Translation.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2013a. Translation of "It" in a Deep Syntax Framework. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*. Omnipress, Inc.

Michal Novák, Zdeněk Žabokrtský, and Anna Nedoluzhko. 2013b. Two Case Studies on Translating Pronouns in a Deep Syntax Framework. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1037–1041. Asian Federation of Natural Language Processing.

Petya Osenova, Rosa Del Gaudio, João Silva, Aljoscha Burchardt, Martin Popel, Gertjan van Noord, Dieke Oele, and Gorka Labaka. 2015. Interim report on the curation of language resources and tools for deep mt. Technical Report Deliverable D2.5, Version 2.0, QTLeap Project.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233. Springer.

Martin Popel, Jaroslava Hlaváčová, Ondřej Bojar, Ondřej Dušek, António Branco, Luís Gomes, João Rodrigues, Andreia Querido João Silva, Nuno Rendeiro, Marisa Campos, Diana Amaral, Eleftherios Avramidis, Aljoscha Burchardt, Maja Popovic, Arle Lommel, Iliana Simova, Nora Aranberri, Gorka Labaka, Gertjan van Noord, Rosa Del Gaudio, Michal Novák, Rudolf Rosa, Aleš Tamchyna, and Jan Hajič. 2015. Report on the first mt pilot and its evaluation. Technical Report Deliverable D2.4, Version 1.8, QTLeap Project.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Comput. Linguist.*, 27(4):521–544.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association.

Olga Uryupina, Alessandro Moschitti, and Massimo Poesio. 2012. BART Goes Multilingual: The UniTN/Essex Submission to the CoNLL-2012 Shared Task. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 122–128. Association for Computational Linguistics.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12. Association for Computational Linguistics.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, page 167–170. Association for Computational Linguistics.