

Pronoun Translation and Prediction with or without Coreference Links

Ngoc Quang Luong
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
nluong@idiap.ch

Lesly Miculicich Werlen*
Université de Neuchâtel
Institut d’Informatique
2000 Neuchâtel, Switzerland
lesly.miculicich@unine.ch

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
apbelis@idiap.ch

Abstract

The Idiap NLP Group has participated in both DiscoMT 2015 sub-tasks: pronoun-focused translation and pronoun prediction. The system for the first sub-task combines two knowledge sources: grammatical constraints from the hypothesized coreference links, and candidate translations from an SMT decoder. The system for the second sub-task avoids hypothesizing a coreference link, and uses instead a large set of source-side and target-side features from the noun phrases surrounding the pronoun to train a pronoun predictor.

1 Introduction

The NLP Group of the Idiap Research Institute participated in both sub-tasks of the DiscoMT 2015 Shared Task: pronoun-focused translation and pronoun prediction (Hardmeier et al., 2015). The first task aimed at evaluating the quality of pronoun translation in the output of a full-fledged machine translation (MT) system, while the second task aimed at restoring hidden pronouns in a high-quality reference translation. In our view, both sub-tasks raise the same question: given the limitations of current anaphora resolution systems, to what extent is it possible to correctly translate pronouns with unreliable knowledge of their antecedents? Although the answer depends on the translation divergencies from the source language to the target one, we explore here two different approaches to answer this question, within the DiscoMT 2015 Shared Task: one using imperfect knowledge of the antecedents of pronouns, and the other one replacing it with a large set of morphological features.

The SMT system we submitted to the pronoun-focused translation sub-task (Section 3) combines

two probabilistic knowledge sources to decide the translation of the English pronouns *it* and *they* into French, namely a probability distribution obtained from an anaphora resolution system and one obtained from the SMT decoder. The classifier for the pronoun prediction sub-task (Section 4), uses morphological and positional features of source-side and target-side noun phrases surrounding the pronoun to be restored, without any hypothesis on its antecedents. System configurations are shown in Section 5, and results in Section 6.

2 Related Work

As rule-based anaphora resolution systems reached their maturity in the 1990s (Mitkov, 2002), several early attempts were made to use these methods for MT, especially in situations when pronominal issues must be addressed specifically such as EN/JP translation (Bond and Ogura, 1998; Nakaiwa and Ikehara, 1995). Following the development of statistical methods for anaphora resolution (Ng, 2010), several studies have attempted to integrate anaphora resolution with statistical MT, as reviewed by Hardmeier (2014, Section 2.3.1). Le Nagard and Koehn (2010) designed a two-pass system for EN/FR MT, first translating all possible antecedents, identifying the antecedents of pronouns using (imperfect) anaphora resolution, and constraining pronoun translation according to the features of the antecedent (with moderate improvements of MT). Other attempts along the same lines include those by Hardmeier and Federico (2010), and by Guillou (2012). Our system for the first sub-task (Section 3) enriches the approach with a probabilistic combination of constraints from anaphora resolution and pronoun candidates from the search graph generated by the MT decoder.

Another line of research attempted to post-edit pronouns in SMT output, possibly including as features the baseline translations of pro-

*Work performed while at the Idiap Research Institute.

nouns. The approach was shown to be successful for translating discourse connectives (Meyer and Popescu-Belis, 2012). A large set of features was used within a deep neural network architecture by Hardmeier (2014, Chapters 7–9). In our system for the second sub-task, we extend the features sketched by Popescu-Belis et al. (2012).

3 Pronoun-Focused Translation

Our system for this task works in two passes. First, the source text is pre-processed and translated by a baseline MT system to acquire pronoun candidates. Then, we apply several post-editing strategies over the translations of “*it*” and “*they*”, which help in correcting erroneous instances.

3.1 Pass 1: Baseline MT Outputs

The test data is first tokenized using the tokenizer provided by the organizers. Then, we apply a baseline MT system to generate the candidate pronouns. This system is the Moses decoder (Koehn et al., 2007) with a translation and a language model trained with no additional resources other than the official data provided by the shared task organizers (including Europarl, News Commentary and Ted talks). Parameters are tuned on domain-specific Ted(dev) data set. We run the Moses decoder with the *-print-alignment-info* and *-output-search-graph* options to obtain the word alignments and the search graph plain-text representation, used for post-editing in the second pass.

3.2 Pass 2: Automatic Pronoun Post-editing

Since the pronoun-focused task concentrates on the quality of translated pronouns, in the second pass we post-edit target words aligned to “*it*” and “*they*” while keeping intact all the others. However, when translating these pronouns into French, the target pronoun is determined not only by the source word itself, but also by other contextual and grammatical factors, and most importantly by the actual gender of the antecedent. Therefore, the whole source sentence and its precedent sentences, as well as the target one, are analyzed for making decision.

3.2.1 Overview of our Approach

Our post-editing process considers the baseline translation of each pronoun “*it*” and “*they*” from the output of Pass 1. If this is one of the “complex” pronouns (e.g. “*celui*” or “*cela*”, see Section 3.2.6), then we simply accept the results from

Pass 1 (baseline translation) and do not attempt to post-edit this pronoun. If this is not the case, then we check first whether it is a subject or an object pronoun. In the former case (subject pronoun), we examine two cues: the gender and number of the translation of its antecedent hypothesized by a co-reference system, along with the decoder’s score for this lexical item calculated from the search graph during decoding. The selected pronoun is the one that maximizes the combined scores of these two criteria. In the latter case (object pronoun), we use a set of heuristics based on French grammar rules to seek the appropriate word. Finally, the post-edited word is substituted to the one from Pass 1 in order to generate the output of Pass 2. These steps are displayed in Figure 1.

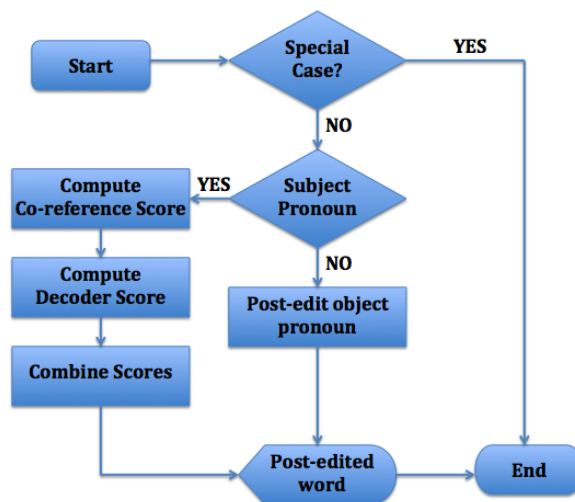


Figure 1: Flowchart of post-editing process

3.2.2 Grammatical Gender and Number

French pronouns always conform to the grammatical gender and number of their antecedent. Ignoring this contextual factor, as current phrase-based MT systems do, may generate inaccurate pronoun translations. Therefore, we consider the antecedent’s gender and number as the most important criterion for pronoun translation.

We thus perform anaphora resolution on the source side, and using alignment we hypothesize the noun phrase antecedent on the target side (French), and determine its gender and number. More specifically, we first employ the Stanford Coreference system (Lee et al., 2011), which currently supports English and Chinese, for identifying the antecedents of the source pronouns (“*it*” or “*they*”). In cases where antecedent is a noun

phrase with several nouns, then the head word is identified by the toolkit using syntactic features extracted from the sentence’s parse tree (Raghunathan et al., 2010). It is very likely that its aligned words will be the target pronoun’s antecedent. A French Part-Of-Speech (POS) tagger is then used (Morfette by Chrupala et al. (2008)) to obtain morphological tags, from which we extract the gender and number of the antecedent.

If the anaphora resolution system always identified accurately the antecedent, then the above method would perfectly post-edit pronouns, with some exceptions: e.g. the case of non-referential pronouns, or antecedents which are singular in form yet plural in meaning (e.g. “*a couple*” ... *they*). However, we estimate that the accuracy of the anaphora resolution system we used was around 60% only, as we found by examining 100 sentences containing 120 pronouns. Therefore, we define a confidence score for coreference resolution based on this accuracy. In other words, if the antecedent detected by the system is masculine singular, then the confidence score for a masculine singular target pronoun is 60%, and for a feminine singular one it is 40%. The decision is made by considering the decoder score presented hereafter.

3.2.3 Decoder Scores

Our motivation for using the decoder score is that the baseline SMT system generates the 1-best hypothesis based on the global feature functions score; however, this does not guarantee that the translations of every word are optimal, especially for pronouns. Hence, we calculate, for each pronoun, the number of occurrences of all its possible translations in the Search Graph (SG) built by Moses during the decoding process.

In the search graph plain-text file (generated by using the *-output-search-graph* option), each line represents a partial hypothesis and stores all its attributes. Among them, we notice two important attributes: “*covered*” (the source word’s position) and “*out*” (the source word’s translation). By selecting the hypotheses whose “*covered*” attribute matches the position of the source pronoun, we can list all possible candidates (in “*out*” attribute) and count the number of occurrences of each type. The decoder score (noted SG), i.e. the probability of translating the source pronoun into a specific target one, is computed as the ratio between its number of occurrences and the sum over all pronoun candidates.

3.2.4 Combination of Scores

We demonstrate the combination of coreference and decoder scores on an example, with the following source text: “*the supreme court has fallen way down from what it used to be .*” and the following MT hypothesis (with several mistakes): “*la cour suprême a chuté de manière ce qu’ il était .*”. Here, the source word “*court*”, detected by the anaphora resolution system as the antecedent of pronoun “*it*”, is aligned to the target word “*cour*”, whose gender and number are determined as feminine and singular respectively. Thus, we consider only two singular candidates “*il*” and “*elle*” as potential translations¹, with the confidence scores computed as above: $p_{ana}(\text{“il”}) = 0.40$, $p_{ana}(\text{“elle”}) = 0.60$. In the next step, the SG enables us to compute the probability to translate “*it*” into either of these candidates, yielding: $p_{SG}(\text{“il”}) = 0.35$, $p_{SG}(\text{“elle”}) = 0.29$. The final scores are simply the averages of the two scores (*ana* and *SG*): $p(\text{“il”}) = 0.375$ and $p(\text{“elle”}) = 0.395$, and the candidate with the highest score (“*elle*” in this case) is selected, leading here to an improved output (in terms of pronoun translation, not overall quality): “*la cour suprême a chuté de manière ce qu’ elle était .*”

3.2.5 Object Pronoun It

In English, “*they*” plays the role of a subject pronoun, since its antecedent is a plural noun phrase. Therefore, its translations into French are generally plural subject pronouns². On the contrary, “*it*” can be used either as a subject or an object. Due to the fact that, unlike English, French singular subject and object pronouns are different, we propose post-editing rules to deal with this case.

Generally, the object pronoun “*it*” refers to the “recipient” of an action caused by the subject, and generally follows the verb. However, its position might be either right after the verb (e.g. “*I know it*”) or several words away (e.g. “*I talk about it*”). In order to detect the object pronouns, we employ Stanford parser (Chen and Manning, 2014). In the parse tree, an object pronoun is always a node of a subtree whose root is a verb phrase (VP) node, while a subject pronoun is under a noun phrase (NP) node. Therefore, we traverse up-ward from

¹All other singular pronouns are considered as special cases, see Section 3.2.6.

²Except when they refer to English plural nouns which are singular in French, e.g. “*trousers*” – > “*pantalon*”.

the pronoun node to the root. If on the way we encounter “VP” node, then we consider the pronoun as an object one.

The translation of “it” depends on the object type (direct or indirect), which we identify by matching the verb preceding the pronoun with one of the French verbs which always have an indirect object³. For direct objects, the translation is *l’* if the following word starts with a vowel or a silent ‘h’, otherwise it is either “*le*” or “*la*” depending on the antecedent’s gender (masculine or feminine, respectively). The SG score is not used for this decision. For indirect objects, the translation is “*lui*”, which is identical for both genders.

3.2.6 Special Cases

We observed on development data that our methods had difficulties with some French pronouns, which require more sophisticated constraints to determine their translation, which the above rules did not fully cover. Indeed, when applying the above rules, the judgments from annotators showed that a large part of these corrections degraded Pass 1’s translation. Therefore we decided not to post-edit the results of baseline SMT (Moses) for: demonstrative pronouns (*ce* or *c’* before a vowel, *ça*, *celui*, *cela*, *celle*, *celui-là* and *celui-ci*); the indefinite pronoun *on*; and two personal pronouns specific to French which have many idiomatic uses (*y* and *en*).

3.2.7 Replacement or Insertion

Due to alignment or translation errors, sometimes a source pronoun is aligned with a non-pronoun target word, which is detrimental for post-editing. Therefore, if the word to be processed is not one of the known French pronouns, we insert the post-edited pronoun in the position preceding it, without replacing the non-pronoun word. For instance, given the following source sentence: “*I see it and then I buy it*” and the Pass 1 (incorrect) hypothesis: “*Je vois et puis j’ achète*”, the MT system aligns wrongly “*see it*” with “*vois*”, and respectively “*buy it*” with “*achète*”. Our post-editing method suggests the following post-editions for the words aligned with “*it*”: “*le*” for the first occurrence, and “*l’*” for the second one. We will not alter the current target words *vois* and *achète*, since they are not known French pronouns. Instead, we add the post-editions in front of them,

³Using the list at <http://instruction2.mtsac.edu/french/li-brete3/chapitre11/verbesobjINdirect.htm>.

yielding the following post-edited target sentence: “*Je le vois et puis j’ l’ achète*”, which has both translations of “*it*” correct.

4 Cross-Lingual Pronoun Prediction

4.1 Training Datasets

The challenge in this task is to build classifiers to predict the hidden pronouns in translations, knowing the source. Four data sets of different domains were provided for development: Europarl, News Commentary (NCv9), IWSLT 2014 and TED(dev) talks. Each data set includes a series of five-element tuples: source sentence, target sentence (with pronouns substituted by placeholders), alignment information, actual pronouns and gold-standard ones (last two not given in the test data).

We first extract features for all occurrences of “*it*” and “*they*”, and then train classifiers over the feature set with various machine learning methods. In fact, to ensure an acceptable training time, we exploit entirely only the smaller data sets, and partially the larger ones: we use for constructing predictors all the occurrences of “*it*” and “*they*” of TED(dev), 10% of those of NCv9, 10% of those of IWSLT and about 1% of those of Europarl. The sizes, total numbers of “*it*” and “*they*” occurrences, and the actual number exploited are shown in Table 1.

Dataset	Size	#(it+they)	#(it+they)
	#sentences	provided	used
NCv9	182761	41227	4123
TED	1664	747	747
IWSLT	179404	77354	7730
EUROPARL	2049662	273827	2700

Table 1: Size, number of occurrences of “*it*” and “*they*”, and instances actually used for training.

4.2 Features

The goal of the submitted system is to explore the potential of morphological features for predicting target pronouns, without attempting to perform anaphora resolution, which is error prone and might not be required, in many cases, for correct pronoun prediction. Instead, we extract possible candidates for antecedents (co-referent nouns and pronouns) from the context surrounding the hidden pronoun and its source counterpart. We aim at estimating how much information we can obtain from the context words without using anaphora

resolution for the prediction. We illustrate the idea on the following pair of sentences as example:

EN: *The police reported the accident to the township, but **it** didn't take action.*

FR: *La police a signalé l'accident à la commune, mais [elle] n'a pris aucune mesure.*

In this case the source pronoun is “it” and the hidden pronoun is “elle”, which must be determined by the system. Two out of the three nouns preceding the hidden pronoun are feminine and singular; therefore, we predict based on the majority gender and number that the pronoun translating “it” into French is singular and feminine, which corresponds to “elle”. In this example we used information of gender and number, but we added also other features that we considered to be potentially relevant.

The features were extracted from both source and target sentences. The target-side features are the 3 nouns or pronouns preceding and the 3 nouns or pronouns following the hidden pronoun. Also, we add as features the gender, number, person, and POS tag for each of these nouns or pronouns. To determine them automatically, we used the French tagger Morfette (Chrupala et al., 2008). Additionally, we included two sets of “summarized” features. The first set corresponds to the modes (i.e. majority) of gender, number and person respectively. For example, if 2 of the 3 preceding nouns or pronouns are feminine, then we indicate that the mode of the gender in the preceding part is *feminine*. Thus, we have 3 modes (gender, number, person) for the preceding nouns/pronouns and 3 for the following ones. The second set of “summarized” features indicates whether all preceding/following nouns and/or pronouns have the same gender, number or person. For example, if all preceding nouns or pronouns are feminine then the value of the feature will be *feminine*, but if only 1 or 2 of them are feminine while the rest are masculine then the value of the feature will be *not-absolute*. Similarly to the first set, we have 3 indicators for the preceding part and 3 for the following part. There are in all 42 features extracted from the French target text.

The 14 source-side features are the original pronoun, the 3 preceding and the 3 following nouns or pronouns, and their respective POS tags identified with the English tagger TreeTagger (Schmid, 1994). Additionally, for each extracted English

noun or pronoun, we included their aligned words in the French text, with the same target-side features as described above (42 features). Finally, we have 98 features to analyze – which represent quite a large set, requiring a large training set for properly learning their relevance.

4.3 Pronoun Prediction

The predictors are trained using the WEKA toolkit (Hall et al., 2009). We experiment with four machine learning techniques: Naive Bayes (NB) (Friedman et al., 1997), Decision Trees (DT) (Quinlan, 1986), Support Vector Machines (SVM) (Burges, 1998), and Random Forests (RF) (Breiman, 2001). With features coming from the four data sets presented above, we train the classifiers and then test them using 10-fold cross validation. For NB, SVM and RF, the default parameters are used. For DT, the “minimum number of instances per leaf” is adjusted from 5 to 15 and binary splits are applied on nominal features. The evaluation results shown in Table 2 indicate that, on all four sets, NB and DT significantly outperform SVM and RF. When comparing between NB and DT, there are cases where the former is more beneficial (e.g. on IWSLT data), but also reverse ones (e.g. on NCv9). Based on these results, we decide to employ Naive Bayes and Decision Trees for our submissions.

Dataset	NB	DT	SVM	RF
NCV9	0.421	0.453	0.401	0.386
TED	0.463	0.476	0.422	0.419
IWSLT	0.560	0.535	0.510	0.498
EUROPARL	0.478	0.466	0.424	0.398

Table 2: Cross-validation results (macro-averaged F-scores) over 4 data sets and 4 types of classifiers.

The size and domain of the data are among the top factors affecting the performance of the classifiers. We prepared three composite data sets from the training data to study these factors:

- **ALL:** all data (large size)
- **IWSLT:** only data from IWSLT 2014 (7703 instances) (in-domain data with the test set)
- **SPL:** sampled data (4123 NCv9 + 7730 IWSLT + 747 TED + 2700 EUROPARL, for a total of 15,300 instances) (partially in-domain data, large size)

These sets are used for training the two most effective machine learning methods found above through cross-validation, namely NB and DT, resulting in a total of six classifiers.

5 Submissions to the Shared Task

5.1 Task 1: Pronoun-Focused Translation

Our submissions were evaluated over a test set of 2093 sentences, containing 1105 pronouns “*it*” and “*they*”, following the above method. In order to better understand the contribution of co-reference information itself to improve pronoun translation, besides the system with combination of two scores as stated above (denoted as **SYS1**), we also submitted another (contrastive) system which only uses the gender of the hypothesized antecedent to correct the subject pronoun (**SYS2**).

5.2 Task 2: Pronoun Prediction

As stated above, the two most effective classifiers were applied to the test set of 2093 sentences, with 1105 instances of “*it*” and “*they*”, yielding predicted labels for each of them. Then, in order to select the two best systems for submission, we sampled a subset of 147 pronouns (“*it*” and “*they*”) and inspected the accuracy of predictions. The two systems with the highest total of accurate instances, namely *DT trained on IWSLT* and *NB trained on ALL*, were selected for submission. Moreover, we observed from these results that using in-domain data for training (i.e. from IWSLT) was more beneficial than using a mixed set. In some cases, the simple NB classifier was more effective than DT on our data.

6 Results and Discussion

The submissions to the first task were judged by human annotators (recruited by the task organizers) for the correctness of translated pronouns, using two main metrics: “Accuracy with OTHER” (all pronouns) and “Accuracy without OTHER” (only on a limited pronoun set). Our system was ranked first, with scores of, respectively, 0.657 and 0.617. Still, these scores remain slightly below the Moses baseline system provided by the organizers (trained on the same data as our system, see Section 3.1). Our scores on the more frequent pronouns (particularly “*il*” and “*elle*”) demonstrate the validity of our approach, while our (still good) scores on the rare ones reflect our strategy to avoid post-editing our baseline SMT output.

Unlike the first task, the strategy we proposed for the second one (using morphological features and no anaphora resolution) obtained rather poor results, ranking among the weakest submissions. Our two submissions scored respectively 20.62 and 16.39 in terms of fine-grained macro-averaged F-score, and respectively 32.40 and 42.53 for coarse accuracy. In fact, as for the first task, the baseline proposed by the organizers (using a language model to restore pronouns) was the best performing strategy (58.40 F-score and 68.42 accuracy). These results tend to show that the proposed features are poor predictors of the pronoun to be used, or possibly that the number of features was too large with respect to the available training data. Using hypotheses from anaphora resolution tends to improve performance, but its contribution remains below the statistical baseline. This indicates the need for additional knowledge, or higher anaphora resolution accuracy, to improve over the baseline.

7 Conclusion and Perspectives

In this paper, we proposed some ideas to enhance the translation quality of pronouns from English into French. For pronoun post-editing (Task 1), coreference scores combined with those from an SMT decoder were employed to correct the wrong pronouns generated by SMT system. Furthermore, with object pronouns, we suggested using specific grammatical rules to determine the candidate. While reaching a high rank compared to other participants, the approach still left a number of pronouns untouched. On the contrary, our rather low scores on Task 2 indicate that unstructured context information is insufficient for predicting pronouns. Therefore, integrating these predictions as an additional feature for the post-editor in Task 1 does not seem promising.

Future work will focus on a deeper analysis of the factors that are most detrimental to current predictors, the selection of co-reference features to train them, and their integration directly into the SMT decoder.

Acknowledgments

The work has been supported by the Swiss National Science Foundation through the MODERN project (www.idiap.ch/project/modern). We are grateful to three anonymous reviewers for their helpful comments.

References

- Francis Bond and Kentaro Ogura. 1998. Reference in Japanese–English machine translation. *Machine Translation*, 13(2):107–134.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- Christopher J. C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.
- Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar.
- Grzegorz Chrupala, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.
- Nir Friedman, Dan Geiger, Moises Goldszmidt, G. Provan, P. Langley, and P. Smyth. 1997. Bayesian network classifiers. In *Machine Learning*, pages 131–163.
- Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of EACL 2012 Student Research Workshop (13th Conference of the European Chapter of the ACL)*, pages 1–10, Avignon, France.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18.
- Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, Paris, France.
- Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, Lisbon, Portugal.
- Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Sweden.
- 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 Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic.
- Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 258–267, Uppsala, Sweden.
- Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford’s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 28–34, Portland, OR.
- 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, France.
- Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London, UK.
- Hiromi Nakaiwa and Satoru Ikehara. 1995. Intrasentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. In *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 96–105, Leuven, Belgium.
- Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1396–1411, Uppsala, Sweden.
- 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 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.
- J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.
- Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 492–501, Cambridge, MA.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.