

Minimally supervised lemmatization scheme induction through bilingual parallel corpora

Taesun Moon and Katrin Erk

Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA

tsmoon, katrin.erk@mail.utexas.edu

Abstract

We present a lemma induction scheme on a target language through minimally supervised alignment and transfer methods utilizing English-to-German parallel corpora. Compared to previous alignment and transfer approaches, the approach outlined here increases computational efficiency and significantly reduces the level of supervision necessary in inducing clusters of inflectional forms. Furthermore, we increase our search field to include not only verbs but also nouns and adjectives in the target language, and achieve comparable results to previous unsupervised monolingual methods.

1 Introduction

Cross-language projection of linguistic information through alignment and transfer methods using parallel corpora has been used for a variety of tasks and purposes such as deriving the syntactic structure of a target language (Wu, 1997), extracting paraphrases (Pang et al., 2003; Bannard and Callison-Burch, 2005), extracting bilingual knowledge (Shin et al., 1996), or semantic disambiguation (Diab, 2000). Among these, one group of approaches has focused on inducing basic NLP tools such as POS taggers, noun chunkers, and morphology analyzers for a given target language (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001; Yarowsky and Ngai, 2001; Drábek and Yarowsky, 2005; Ozdowska, 2006; Moon and Baldrige, 2007). The latter approaches take note of the fact that many of the

languages of the world are underdocumented and resource challenged, and that there is a need to provide rudimentary but robust tools to assist in the process of documentation and analysis.

The study outlined here is in line with this basic premise. Using sentence-aligned parallel texts from the Europarl Corpus (Koehn, 2005), with English as the source and German as the target, we induce a lemmatization scheme over the target language through alignment and transfer methods. The same problem has been dealt with only once before in Yarowsky et al. (2001), which treated verbs in Czech and French. We instead propose a different approach which foregoes some of their basic assumptions as well as a probabilistic model based on those assumptions and instead focuses on methods for reducing the search space of candidate lemmata, and expands the lemmatization to incorporate not just verbs but also nouns and adjectives. With an overall token precision of 0.836, we achieve results comparable to other unsupervised methods. Given that the induction of lemmatization schemes is an important stepping stone in building other fundamental NLP tools such as lemmatizers, POS taggers, and parsers, we believe the aims and results of this study can provide useful insights as well as critical data in this process of accumulating tool sets.

In this paper, after a review of related work including a discussion of the models and assumptions in Yarowsky et al., and a presentation of the data sets and programs we will be using, we outline and present our own approach which, despite failing to best the results of Yarowsky et al., show that robust results are possible in spite of a significantly

reduced level of supervision, even when the target word categories are expanded to include not only verbs but also nouns and adjectives. In section 5, we provide the results of our attempt to reimplement Yarowsky et al., present our own results, and evaluate them according to two criteria, one of which is novel in its assessment of hard clustering tasks such as the lemmatization task attempted here and is more methodologically sound given the nature of the task.

2 Related Work

Unsupervised monolingual morphology segmentation is a topic that has been tackled many times in the literature (Goldsmith, 2001; Sassano, 2001; Goldwater, 2006; Hammarström, 2006; Creutz and Lagus, 2007). Though such approaches generally manage to provide relatively reliable segmentation schemes with precisions between the ranges of 0.8 and 0.9, it is difficult to generalize beyond the segmentation of individual word types to how they relate to the POS categories in a given language or its syntax. We show in this paper that alignment and transfer methods based on utilizing the linguistic metadata of a well-documented language for the analysis of another can provide concrete motivation for limiting the search space for potential clusterings of inflected forms and can also impose higher-level syntactic constraints. This will result in an analysis that is less dependent on the quirks of orthographic similarities.

Yarowsky et al.(2001) introduced the method of lemmatization scheme induction through alignment and transfer methods. It forms part of a larger group of studies that focus on the use of bilingual corpora to induce NLP tools for a target language (Shin et al., 1996; Wu, 1997; Diab, 2000; Yarowsky et al., 2001; Drábek and Yarowsky, 2005; Ozdowska, 2006). In this study, a core algorithm in the induction of lemmatization schemes in a target language is the transitivity function, an approach based on the intuition that if one lexeme and another lexeme in the target language have been aligned with more lemmas in the source language than with some other lemma, the more likely it is that the two words can be grouped together under some meaningful cluster. They use the following probabilistic model:

$$P(T_{lemma}|T_{infl}) = \sum_i P(T_{lemma}|S_{lemma_i})P(S_{lemma_i}|T_{infl}) \quad (1)$$

In this approach, the probability that a target lemma T_{lemma} will be the lemma of an inflected token in the target T_{infl} is estimated by summing over the probability of T_{lemma} given a lemma S_{lemma_i} in the source multiplied by the probability of the source lemma given T_{infl} for all the lemmas in the source. The transitive links used here will increase the likelihood of $P(T_{lemma}|T_{infl})$ the more often they occur over all source lemmas which provide a link between the two.

The major limitation of this approach is that it requires a pre-selected list of lemmata in the target language. Though it is possible to modify the model and implementation so that no assumptions are necessary regarding which word types in the target are lemmata, or “dictionary entry forms”, and which are inflected forms, such a modification comes at the cost of considerable time complexity. Needless to say, manually selecting a set of target lemmata as has been done in this study is a step which significantly increases the level of supervision.

A second limitation of this approach (discussed with examples in Section 5) is that the transitivity function when implemented without any assumptions regarding lemmata casts a very wide net, favoring retrieval over precision. Even when implemented on a manually selected set of target lemmata, Yarowsky et al. impose a empirically determined threshold (which is unspecified) on the transitivity function to limit the size of the sets of candidate inflectional forms which have been associated with some candidate target lemma. It is not discussed whether this same threshold is applicable across a wide spectrum of languages, and further investigation might reveal that a case-by-case inspection of the data is required in each instance to determine this threshold.

With this approach, they post a precision of 0.992 and a recall of 0.994 over word tokens for the 12M word French Hansards using the alignment method alone. However, it should be noted that the induction was performed for only verbs in the target language and that the study had implemented a POS tagger induced through similar minimally supervised means.

They gain a small increase in precision and a substantial increase in retrieval over target word types by augmenting the above approach with a trie based search and a backoff model based on Levenshtein distance and distributional similarity measure. With the aid of these methods, they increase the general level of precision over all their target corpora to an almost insuperable 0.99 and a retrieval of 1.00. The latter augmentative approaches are outlined in more detail in Yarowsky and Wicentowsky (2000).

3 Data and Alignment

3.1 Europarl Parallel Corpus

The German and English sections of the Europarl parallel corpus (Koehn, 2005) were used in this study. The Europarl parallel corpus is a collection of texts in 11 languages extracted from the proceedings of the European parliament with each text comprising some 25 to 30 million words. Any two texts from this corpus are mutually parallel.

To enhance the accuracy of the parsing and alignment tasks, the parallel corpus was further trimmed to English sentences of less than 45 words in length. This reduced the size of the English and German corpus to roughly 17 million words each.

3.2 Lemmatization and POS tagging of source text

POS tagging for the English text was done with the maximum entropy based C&C tagger (Curran and Clark, 2003), which was trained on the Wall Street Journal of the Penn Treebank. The POS tagged source text was then supplied to the lemmatizer, Morpha (Minnen et al., 2001), a finite state morphology analyzer, whose only requirement for prior POS tagged data is that verbal tags are headed by a V and noun tags other than proper nouns are headed by an N. Such knowledge of the word category of a lexeme is necessary in enhancing the performance of Morpha.

3.3 Word Alignment

Word alignment between the two texts was achieved with GIZA++ (Och and Ney, 2003). The alignment was made with English as the source and German as the target. In this stage, the parallel corpus was further reduced to three-quarters of the trimmed corpus

derived in the stage outlined above. It was a process recommended in Yarowsky et al. (2001) to reduce undue noise in the alignment model, and so alignments with a confidence measure in the lower 25% of the parallel corpus were removed from consideration for this study.

3.4 TIGER Treebank Corpus

The TIGER Treebank (Brants and Hansen, 2002) corpus was used as the evaluation corpus on which to test lemmatization schemes. The corpus, which is currently at version 2.1, is a collection of German newspaper text gathered from the Frankfurter Rundschau and consists of app. 900,000 tokens. It is annotated with POS tags and lemmata for terminal nodes and has been manually annotated for syntactic information. The use of this corpus also allowed us to evaluate how well the scheme induced from one domain would translate to another.

4 Approach

Our approach wholly does away with the transitivity function by aggressively culling the search space for candidate lemmata and candidate inflected forms. First, we limit the set of candidate lemmata to the word types in the target language which have the greatest possibility of being associated with some lemma in the source language. With this candidate lemma, we generate one set of lemmata to inflected form mappings by limiting the linkages to those source lemmata and target word type associations which exceed a manually determined probability threshold. We generate a second set of mappings from a candidate lemma to a set of target word types which has been limited to those which have been observed in alignment with a source lemma and then further reduced through an automatically induced edit distance threshold.

4.1 Lemmatization candidate trimming

Using the word based alignment output from GIZA++, we obtained the conditional likelihood estimates from the target text:

$$P(\ell_s T_s | w_t) \quad (2)$$

$$P(w_t | \ell_s T_s) \quad (3)$$

where subscripts s and t are source and target texts, respectively, ℓ and T are lemma and POS tag, respectively, and w is a word type in the target language. ℓ_s is an element of the set Λ_s which is the set of all lemmata observed in the source language and w_t is an element of the set W_t which is the set of all types observed in the target language. In comparison to the two previous attempts to lemmatize a target language through alignment and transfer methods (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001), we expand the set of POS tags from verbs to incorporate adjectives and nouns as well, in short all the content word categories in English except for the adverbs.

Therefore, all lemmata in the English source had to be considered with their respective POS tags, considering that many lemmata in English can be ambiguous with regard to word category when judged on their surface form alone. Hereinafter, to simplify notation, all source lemma arguments in functions shall be assumed to also be tagged with relevant POS information. As such, the above equations are equivalent to

$$P(\ell_s|w_t) \quad (4)$$

$$P(w_t|\ell_s) \quad (5)$$

Also, note that words in the aligned target text are merely assumed to be a word type in the most general sense, since no assumptions can be made at this point whether a particular word form observed in the target language is the inflected form of some lemma or is itself the general “dictionary entry form”.

In the estimation of the probabilities in (4) and (5), we make an unjustified but practical decision to limit the set of target word types under examination to those which have string lengths of four or longer. This was mainly due to the fact that the Levenshtein edit distance algorithm is incapable of calculating meaningful scores when the strings being compared are both very short.

To limit the search space, we build two mapping tables, one from the target word types to the source lemmata and another from the source lemmata to the target word types.

The mapping from the target to the source, $TS : W_t \rightarrow \Lambda_s$, is built by

$$TS(w_t) = \ell_s \text{ iff } P(\ell_s|w_t) > 0.75 \quad (6)$$

The mapping from the source to the target, $ST : \Lambda_s \rightarrow W_t$ is built by

$$ST(\ell_s) = \arg \max_{w_t} P(w_t|\ell_s) \quad (7)$$

The mapping from target to source TS is ensured to be unambiguous since the probability threshold for assigning a mapping from w_t to ℓ_s is 0.75. This threshold value was selected after an initial examination of the data extracted from the Europarl corpus; and given its high threshold, it is expected to generate high confidence candidates regardless of the target language. However, future studies will need to examine methods of automating the threshold extraction procedure.

Using the two mappings TS and ST , we will automatically determine a minimal Levenshtein edit distance threshold by comparing the edit distance between all possible W_t to W_t mappings,

$$ST(TS(w_t)) = w'_t \quad (8)$$

where $w_t, w'_t \in W_t$. The mapping obtained here will be necessary for limiting the search space for the first set of candidate lemma to candidate inflectional form mappings.

```

Declare:  $a[0 \dots n]$ 
1: for  $j$  from 0 to  $n$  do
2:    $a[j] := 0$ 
3: end for
4: for all  $w_t \in W_t$  do
5:   if  $TS(w_t) \neq \text{NONE}$  then
6:      $w'_t := ST(TS(w_t))$ 
7:      $d := \text{edit\_distance}(w_t, w'_t)$ 
8:     if  $d < n + 1$  then
9:        $a[d] := a[d] + 1$ 
10:    end if
11:  end if
12: end for
13: return  $\min(a[0 \dots n])$ 

```

Figure 1: Algorithm for computing edit distance threshold

The specific algorithm for computing the edit distance threshold is laid out in Figure 1. We obtain the edit distance for every w_t, w'_t pair in (8), and keep count of how many times each edit distance score

was observed (which is stored in an array a of length n in the algorithm; in this case, we used an array of length 9). Finally, the edit distance threshold is determined to be the minima among the frequency counts by edit distance score. The actual frequencies can be observed in Figure 2, the graph of which approximates a convex function. Furthermore, even if the number of edit distance scores we keep track of is increased to include all edit distance scores, it is evident that a score and its frequency count will continue to increase until reaching some asymptotic upper limit for all real-word data. Therefore, though the highest edit distance score we maintain a frequency count of is 9, there is no possibility that the frequency count will decrease at some point above that score. The intuition behind the approach is that two target words which have an edit distance beyond a certain threshold is more likely to be noise and those which do not exceed it will be related within some inflectional paradigm; and that this threshold exists at the minima of the frequency counts.

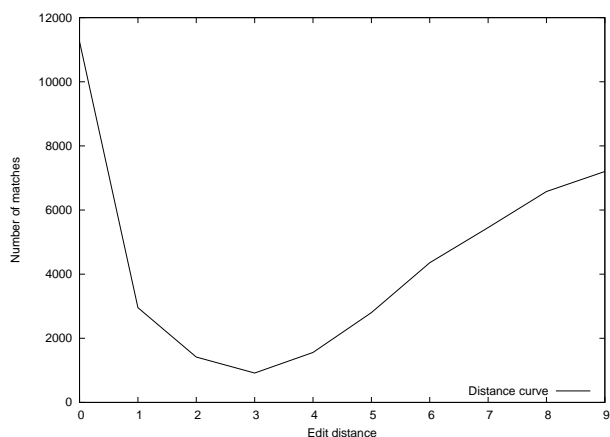


Figure 2: Extraction of Levenshtein edit distance threshold

4.2 Candidate set induction

We induce our first set of lemma group candidates as follows. First, we generate a mapping M from a source lemma ℓ_s to a set of target word types $\Omega_t \subset W_t$ where

$$\Omega_t = \{w_t | P(w_t | \ell_s) > 0\}$$

With this mapping

$$M(\ell_s) = \Omega_t$$

we further trim Ω_t by pegging the lemma candidate as $ST(\ell_s)$ (see equation (7)) and removing all the elements in Ω_t which have a Levenshtein edit distance score from $ST(\ell_s)$ greater than the distance threshold 3 (obtained through the algorithm in Figure 1), resulting in Ω'_t , a subset of Ω_t .

Thus, we have obtained a set of lemma candidates Λ_t in the target language

$$\Lambda_t = \{\ell_t | \forall \ell_s \in \Lambda_s, ST(\ell_s) = \ell_t\}$$

and a set of inflections associated with each ℓ_t in Λ_t

$$C_1(\ell_t) = \Omega'_t$$

Furthermore, the candidate lemma ℓ_t inherits the POS tag from the source language, so that ℓ_t is also specified for whether it is an adjective, noun, or verb.

A second candidate set, or a mapping from candidate lemma to candidate inflected forms, is induced by trimming the mapping TS to a subset of mappings where if the length of the common substring between the input and the output is less than 4, it is removed. However, the common substring in this case is not the longest common substring assumed in general, but merely the common substring from the beginning of each string being compared.

The justification for this is as follows. A very simple assumption can be made that a language will be either prefixal or suffixal in its inflectional system. By implementing two tries over the entire set of word types in the target language W_t , one trie starting from the beginning of the strings and another starting from the end of the strings¹, we can compare how many terminal nodes there are for the forward trie and the reverse trie, the intuition being that the more terminal nodes a particular trie has, the less likely it is that morphological affixation occurs at the terminal nodes of that trie. In the case of our study, it was found that the forward trie had 898 terminal nodes whereas the reverse trie had 4387 terminal nodes. Hence, we come to the simplified conclusion that the target language was suffixal rather than prefixal in generating inflected forms.

The second candidate lemma to candidate inflection mapping, unlike the first candidate mapping, is not from a word type to a set, but from a word type

¹Again, the word types that were submitted to the trie were restricted to those whose length was greater than 3

to a word type. We define the second candidate mapping C_2 as follows:

```

1: for all  $w_t \in W_t$  do
2:   if  $ST(TS(w_t)) \neq \text{NONE}$  then
3:      $w'_t := ST(TS(w_t))$ 
4:     if  $CS(w_t, w'_t) < 3$  then
5:        $C_2(w_t) = w'_t$ 
6:     end if
7:   end if
8: end for

```

where CS is a function on two strings which returns an integer value of the longest common substring starting from the beginning of the two arguments and $ST(TS(w_t))$ is the mapping stated in (8).

Finally, we combine the two candidate mappings into a final candidate mapping C which is a relation from a word type to a set of word types. If there are coinciding ℓ_t in C_1 and C_2 , then the output of C_2 is merged into the set generated by C_1 . Otherwise, candidates are simply added to the mapping C .

5 Results and Evaluation

5.1 An examination of the transitivity function

In our implementation of the transitivity function in (1), we modified the model so that it would not make any assumptions about which words in the target are lemmata and which are not. However, this revealed itself to be computationally too intense in terms of time complexity. When we limited the candidate set of target text lexemes to about 100 and the set of source text lemmata to 50000, it took 60 minutes to complete the computation. It would have been impossible to expand the set of target text lexemes to the 50000 word types that we had. When we reduced the number of source text lemmata to a manageable 1000 words, we were confronted with the problem of sparse data and the function was not able to properly link candidate lexemes. Finally, we tried the option of reducing the set of source lemmata and target lexemes to those which had been observed in transitive verb/direct object relations in the source, the syntactic relations of which were obtained through the C&C tagger (Curran and Clark, 2003).

A small subsample of the results can be observed in Figure 3. In addition to the examples observed in the subsample, the amount of noise in the results in general were excessive and ultimately unfit for in-

ducing lemmatization schemes. In addition to manually defining a set of target lemmata, Yarowsky et al. (2001) used a manually set threshold for the transitivity values obtained through Equation 1 to remove the unfit pairings between candidate lemma and candidate inflected form. While such a threshold over this data might have reduced the level of noise, in the end, it would have been prohibitively time consuming to achieve an enhancement in retrieval or precision over our data set.

5.2 Revised approach

There were 193582 word types in the German portion of the Europarl corpus. From this set W_t , 15945 lemma candidates were induced after applying the culling outlined in section 4. These lemma candidates were mapped to a total of 29056 candidate inflected forms, an average of 1.8 inflectional candidates to a lemma candidate.

Evaluation was conducted using two separate measures. One was over the tokens observed in the TIGER corpus (Figure 4) and another was over types (Figure 5).

	ADJ	N	V	OVERALL
Precision	0.711	0.903	0.718	0.836
Recall	0.277	0.330	0.080	0.267
F-Score	0.399	0.483	0.144	0.405

Figure 4: Scores by tokens and POS tag

	ADJ	N	V	OVERALL
Precision	0.711	0.795	0.840	0.772
Recall	0.822	0.899	0.463	0.874
F-Score	0.762	0.844	0.596	0.791

Figure 5: Scores by types and POS tag

To evaluate type accuracy, we use a measure similar to the Jaccard distance between true and induced inflectional forms for a lemma. Unlike problems of soft clustering, it is possible to define what is a correct clustering in a lemmatization problem. The precision of an individual clustering can be defined as the size of the intersection between an induced set of inflectional forms and the standard set of inflectional forms divided by the size of the standard set.

	LEMMA	INFLECTIONS
VERBS	<i>ergänzen</i>	unternommenen betriebsrat ergänzung abrunden abkehr zusammenführen vervollständigen weswegen flüchtlingskonvention entwicklungschancen staatsangehörigkeit ergänzend ergänzen einander durchschlagen . . .
	<i>sterben</i>	sterben verhungern helfen designierten jährlich zutritt meistens amerikanischen irakern fünfte tod planeten industriegebieten fonds dramatisch us-regierung
NOUNS	<i>knie</i>	asiatischen zusammengestellt zufügt kniefall knie knien apartheid-regime rechtsanspruch
	<i>euro</i>	ausübt euroraums euroumstellung euro-raums euros euro-länder euro-ländern euro-zusammenarbeit euro euro-raum euroländer euro-system . . .

Figure 3: A list of German candidate verb and noun lemmas and their inflected forms extracted automatically through alignment and transitive linkage. List of candidate inflections is unordered either in terms of frequency or in terms of dictionary precedence.

By summing the individual clustering precision figures over the entire set Λ of sets of inflectional forms I_i , and normalizing this by $N = |\Lambda|$ the precision is calculated as

$$\frac{1}{N} \sum_{I_i \in \Lambda} \frac{|I_i \cap I_g|}{|I_i|}$$

Similarly, recall is defined similar to the above but divided by $|I_g|$ instead:

$$\frac{1}{N} \sum_{I_i \in \Lambda} \frac{|I_i \cap I_g|}{|I_g|}$$

These results are given in Fig. 5.

6 Conclusion

We have outlined a minimally supervised approach to inducing a lemmatization scheme for a target language using alignment and transfer methods across parallel bilingual corpora. Compared to the few previous studies on lemmatization (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001), we have reduced the level of supervision necessary to a bare minimum, obviating any need to manually select a set of “dictionary entry forms” for the target language, while retaining a time complexity that is feasible in spite of a lack of predefined assumptions, even when the parallel corpora span some 25M words for both source and target language with app. 200K word types in the target text.

In future studies, to further increase the robustness and accuracy of the approach, several avenues of investigation will have to be included. First, given that it is possible to automatically generate a POS tagger (its robustness and accuracy notwithstanding) for a target language through alignment and transfer methods, it should be possible to leverage such additional information to enhance the accuracy and coverage of our lemmatization method. Second, given current developments in the field, it would be possible to generalize over the induced lemmata set to generate new inflections. To do so would require induction of abstract inflectional patterns in the target language for what may or may not be equivalent or analogous to number, case, tense, mood, voice, etc. which would require the incorporation of all the lemmata over all POS tags observed in English (e.g. prepositions, pronouns, conjunctions, etc.) as well as the syntactic information generated by parsers.

Acknowledgements

This work was supported by NSF grant BCS-0651988. The authors would also like to thank Jason Baldrige and Alexis Palmer for providing invaluable comments on the paper.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604,

- Morristown, NJ, USA. Association for Computational Linguistics.
- Sabine Brants and Silvia Hansen. 2002. Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pages 1643–1649, Las Palmas.
- Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3.
- James R Curran and Stephen Clark. 2003. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*.
- Mona Diab. 2000. An unsupervised method for multilingual word sense tagging using parallel corpora: a preliminary investigation. In *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.
- Elliott Franco Drábek and David Yarowsky. 2005. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 49–56, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.
- Sharon Goldwater. 2006. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.
- Harald Hammarström. 2006. A naive theory of affixation and an algorithm for extraction. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 79–88, New York City, USA, June. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.
- Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Nat. Lang. Eng.*, 7(3):207–223.
- Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for middle English through alignment and projection of parallel diachronic texts. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 390–399.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Sylwia Ozdowska. 2006. Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora. In *EACL 2006 Workshop on Cross-Language Knowledge Induction*.
- Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 102–109, Morristown, NJ, USA. Association for Computational Linguistics.
- Manabu Sassano. 2001. An empirical study of active learning with support vector machines for Japanese word segmentation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 505–512, Morristown, NJ, USA. Association for Computational Linguistics.
- Jung H. Shin, Young S. Han, and Key-Sun Choi. 1996. Bilingual knowledge acquisition from Korean-English parallel corpus using alignment method: Korean-English alignment at word and phrase level. In *Proceedings of the 16th conference on Computational linguistics*, pages 230–235, Morristown, NJ, USA. Association for Computational Linguistics.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.
- David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.
- David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 207–216, Morristown, NJ, USA. Association for Computational Linguistics.
- David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.