Polish-English Word Alignment. Preliminary Study

Alina Wróblewska

Institute of Computer Science, Polish Academy of Sciences,Warsaw, Poland

Abstract. As word alignment is an important topic in statistical machine translation domain, bilingual dictionary extraction or linguistic information projection studies, a lot of attention has been dedicated to improve its quality. However not all languages are sufficiently represented in these examinations. In the following, we give a description of experiments with the Polish-English word alignment training. The aim of our preliminary study on this topic is to identify training parameters of the optimal Polish-English word alignment and to improve training methods themselves. The quality of the Polish-English word alignment is evaluated against a manually created gold standard corpus.

1 Introduction

The idea of the word alignment[1] comes from the large field of the statistical machine translation. The concept of word alignment has been introduced by [5] as an intermediate result of the statistical machine translation. According to them, information about alignments between foreign and target words is a hidden variable in the statistical translation model. Hidden alignments are iterative revealed with the *expectation maximisation algorithm* (EM-algorithm), which estimates parameters of the statistical translation model.

The statistical machine translation concept as described in [5] or [10] is referred to as the *estimation approach* compared to the *association approach* applying heuristic models. In a heuristic model, alignments with the highest association scores between words are selected. The association score is calculated with one of the association measures such as *t-score* [2] or $\textit{log-likelihood}_{ratio}$ [14]. We will follow the estimation approach, in order to identify the optimal Polish-English word alignment.

The main goal in our experiment consists in training and evaluating Polish-English word alignments. According to our knowledge, there has not been published any case study investigating this topic. In our work, we would like to verify existing methods of automatic identification of alignment links. We will test different configurations of IBM models, looking for the best one that outputs the optimal word alignment. Then, we will validate different symmetrisation methods and other concepts of improving the word alignment quality.

[1] We translate the term *word alignment* into Polish as 'przyporządkowania słowne'.

Word alignment may be used not only in the statistical machine translation, but also in extraction of bilingual dictionaries [14], projection of linguistic information [16], etc. In all cases, results depend on the word alignment quality. As our future work will consist in making use of automatically generated alignment links for projection purposes, word alignments need to be as accurate as possible. We intend to find out, to what extent the quality of projected information depends on the quality of word alignments. Results presented in this paper will constitute our point of reference, while evaluating the quality of projected linguistic information.

Our paper is structured as follows. Section 2 introduces the idea of the statistical word alignment and focuses on methods of improving automatic word alignment. Section 3 presents our training and test data. In section 4 we conduct experiments, discuss results and compare them with related works. Section 4 concludes.

2 Statistical Word Alignment

The bilingual word alignment is a mapping between a foreign sentence $F = f_1, ..., f_j, ..., f_t$ and its target translation $E = e_1, ..., e_i, ..., e_s$. In its ideal form, the word alignment is understood as annotation of minimal translational correspondences. An alignment link $al(i, j)$ corresponds to the equivalent translations between e_i and f_j. The set of all alignment links AL is a subset of the Cartesian product $al \in AL : E \times F^2$.

Besides manual annotation of parallel data by bilingual experts, the word alignment for each sentence pair in the parallel corpus may be generated automatically with statistical alignment models such as the so-called IBM [5] and Hidden-Markov [15] models. Mentioned models are implemented in GIZA++ [10], which is a language independent and widely used open source toolkit. Each model is an implementation of two steps of the EM-algorithm. In the expectation step the model is applied to data and missing alignment links between foreign and target words are filled with the most likely values. In the maximisation step, counts for word translation over all weighted alignments are collected and the new translation probability distribution is estimated given these counts. The model parameters are updated and the new model may be used for the next iteration. IBM models are running successively, i.e., the output of a current model training is given as an input to a higher-order model training. We will refer to the order of executing models and the number of model iterations as model training configuration or model training scheme. The translation model training outputs the most probable alignment (Viterbi alignment) of the last iteration.

While translating, a translation model generates one or more words in a target language from an aligned word in a foreign language. Each target word (e_j) is aligned to one $(f_{a(j)})$ or no foreign word (the $NULL$ token). We may

[2] $E \times F = \{(e, f)|e \in E \text{ and } f \in F\}$

differentiate the following alignment scenarios: `one(many)-to-one` (at least one target word maps to one foreign word), `one(many)-to-NULL` (at least one target word maps to the NULL token treated as a foreign word), and finally some source words remain unaligned. As an alignment relation is not only directed, but also it is regarded as a function, the reverse direction is not possible and we cannot map one target word to multiple foreign words. However, `one-to-many` or `many-to-many` alignments between languages are common and it is a major limitation of IBM models that they are not able to generate such links.

2.1 Symmetrisation

Since their conceptualisation [5], IBM models have been extended and the word alignment induced with these models has improved. As we previously mentioned, the statistical word alignment can only handle `one-to-one` and `many-to-one` links. In order to improve the word alignment quality and overcome limitation of alignment scenarios, a method of bidirectional alignments symmetrisation has been proposed by [10]. Bidirectional word alignments (foreign-to-target and target-to-foreign language) derived from the IBM models training are combined together according to a symmetrisation heuristic ([10], [8]). Applying of symmetrisation as a post-processing step of model training improves the quality of the word alignment by providing one-to-many or many-to-many alignment links.

The most fundamental symmetrisation methods are union and intersection [10]. In the union, all alignment points that occur in either of the bidirectional word alignments are unified ($A_{UNION} = A_{S \to T} \cup A_{T \to S}$). The intersection set contains all alignment points coexisting in both bidirectional word alignments ($A_{INTERSECTION} = A_{S \to T} \cap A_{T \to S}$). The union set is supposed to be characterised by high recall and low precision scores. The intersection, in turn, achieves high precision and low recall.

Next to union and intersection, there exist some more complex symmetrisation concepts. One of them is the `grow-diag-final` method presented in [8]. The `grow-diag-final` algorithm starts the symmetrisation process with the selection of all intersection links, which are quite reliable as they are characterised by high precision values. Next, the most reliable candidate alignment points from the union set are added in successive steps. In the first `grow-diag` step, the neighbouring (also diagonally) alignment points between words of which one is unaligned are added to the final set of alignments. In the `final` step, not neighbouring alignment points between words of which one is unaligned are added to the final word alignment. Optionally, the `and` step adds not neighbouring alignment points linking unaligned words. The growing method that combines bidirectional word alignments derived from the IBM model training has been implemented in the statistical machine translation system Moses [9]. We will use this tool in our experiments to perform symmetrisation.

2.2 Improving Word Alignment

The word alignment underlies the statistical machine translation, which is one of the most important and deeply explored topics in the language processing domain. Thereby, a lot of efforts have been made to improve the quality of the word alignment. Some ideas consist in reduction of the vocabulary size, the word alignment is trained on, with stemming [6] or lemmatising [4] techniques. Both related works consider the inflecting – isolating language pairs, i.e., Romanian – English [6] and Czech – English [4] and they report improvement of the word alignment quality. Following this, we are going to make use of lemmatisation and test if training the Polish-English word alignment on lemmatised data enhances the accuracy.

3 Data

3.1 Training Set

We use a Polish-English parallel corpus consisting of 902,874 sentence pairs extracted from sources available for both languages. The main part of our corpus consists of legislative texts of the European Union (JRC-Acquis Multilingual Parallel Corpus [12]). The rest of the corpus is collected from the open source parallel corpus – OPUS [13]: European Medicines Agency documents and documentations (KDE, KDE4, PHP). We take over sentence pairs as they are. As the process of sentence aligning was automatically performed without any manual checking in both cases, corpora may contain misaligned sentences or segments. After removing all duplicates, we get 902,874 sentence pairs. First two columns in Table 4 give the parallel corpus size in tokens and the average number of tokens per sentence.

3.2 Gold Standard

The quality of the word alignment is evaluated against a manually annotated gold standard that has been created for purposes of presented experiments. The gold standard corpus used in our experiment consists of 100 sentence pairs randomly selected from the entire parallel corpus. The selected sentence pairs are manually annotated by three annotators (native speaker of Polish proficient in English). The annotators follow guidelines solving some problematic issues (e.g., annotation of articles, case markers, dropped pronouns, reflexive markers, dates, subordinate clauses). The manually produced annotations are unified according to the following procedure. If at least two annotators assigned the same link, it has been determined as a gold/sure alignment link. We do not differentiate between sure and possible alignment links. We will evaluate the word alignment accuracy against this gold standard corpus.

4 Experiments and Results

4.1 Evaluation Metrics

We evaluate word alignments with the following metrics: *precision, recall, alignment error rate (AER)* and $F_\alpha - measure$ [7].

$$precision(A,S)^3 = \frac{|S \cap A|}{|A|} \qquad recall(A,S) = \frac{|S \cap A|}{|S|}$$

$$AER(A,S) = 1 - \frac{2|A \cap S|}{|A|+|S|} \qquad F_\alpha{-}measure(A,S,\alpha) = \frac{1}{\frac{\alpha}{precision(A,S)} + \frac{(1-\alpha)}{recall(A,S)}}$$

The alignment error rate (AER) is commonly used to evaluate word alignments. Showing the rate of alignment errors, it lays the foundation for improvements of the word alignment quality. However, [7] has criticised the AER metric as not showing any correlation between the measured alignment quality and the statistical machine translation performance. Furthermore, AER does not penalise varying precision and recall, when a distinction between sure and possible links is made. In spite of this criticism, we measure AER of Polish-English alignments in order to compare our results with the outcome of experiments with Czech-English[4] language pair [3], measured with the AER metric.

4.2 Experiment 1: Schemes of the IBM Models Training

In the first experiment, we test different training configurations of the IBM models 1-4[5] [5] and the Hidden Markov model (HMM) [15]. The experiment is conducted on tokenised and lowercased[6] surface word forms in parallel sentences. Evaluated model training configurations and results[7] are shown in Table 1.
We achieve the best word alignment quality (76.8% F_α-measure) for the Polish-English language pair trained with the following IBM models training scheme:

[3] Explanation:
A – automatic alignment links
S – gold standard alignments
α – parameter determining a trade-off between precision and recall

[4] As previously mentioned, no studies concerning Polish-English word alignment have been published. However, the Polish morpho-syntactic structure is similar as in case of the Czech language, which is one of the best-studied Slavic languages. Therefore, we use Czech word alignment as our point of reference.

[5] Even if the remaining IBM models 5 and 6 are sophisticated, they do not significantly improve the alignment quality. Furthermore, training of these models is very time consuming. That is why we limit the current training procedure to the IBM models 1-4 and HMM.

[6] Tokenising and lowercasing are carried out with the standard scripts provided in the Moses toolkit.

[7] Bidirectional word alignments are trained on tokenised and lowercased surface forms of parallel data and symmetrised with the `grow-diag-final-and` algorithm.

Model configuration	Precision	Recall	F$_{\alpha}$-measure	AER
I(1)-HM(1)-III(1)-IV(1)	70.0	62.2	65.9	34.1
I(5)-HM(5)-III(1)-IV(1)	75.2	68.6	72.1	27.9
I(5)-HM(5)-III(3)-IV(3)	77.4	74.5	75.9	24.1
I(5)-II(5)-HM(5)-III(3)-IV(3)	76.7	73.9	75.3	24.7
I(5)-HM(5)-III(5)-IV(5)	76.8	75.2	76.0	24.0
I(10)-HM(10)-III(3)-IV(3)	78.2	74.6	76.4	23.6
I(10)-HM(10)-III(5)-IV(5)	77.7	**75.8**	**76.8**	**23.2**
I(20)-HM(20)-III(3)-IV(3)	**78.8**	74.3	76.5	**23.2**

Table 1. Results. Training configurations of HMM and IBM models. Abbreviation convention: **Roman numerals** stand for the IBM model number; **HM** is an abbreviation of the Hidden Markov model; **Arabic numeral** in brackets stands for the iterations number of the current model.

model I (10 iterations), HMM (10 iterations), model III (5 iterations), model IV (5 iterations), i.e., I(10)-HM(10)-III(5)-IV(5) according to our abbreviation convention. We compare configuration of the IBM model training regarding F$_{\alpha}$-scores. However, it is worth noting that the best configuration gains not only the highest F$_{\alpha}$-score, but also obtains quite balanced precision and recall values. Concerning the slightly worse result achieved by the second best IBM model configuration with the doubled number of iterations of IBM model 1 and HMM (I(20)-HM(20)-III(3)-IV(3)), we suppose to deal with the over-fitting problem. That is why, we decide to rely on the best IBM model configuration (I(10)-HM(10)-III(5)-IV(5)) in the following experiments, without any potential over-fitting threat. We consider the word alignment trained on the best IBM model configuration as our baseline.

4.3 Experiment 2: Symmetrisation

In the second experiment we verify if the `grow-diag-final-and` symmetrisation heuristic is the best one for unifying bidirectional Polish-English word alignments. We compare the output by `grow-diag-final-and` algorithm with results of symmetrising word alignments with `intersection`, `union` and `grow-diag-final` methods.

According to our assumptions, the `grow-diag-final-and` algorithm achieves the best results (76.8 F$_{\alpha}$-measure) compared to other symmetrisation methods. The intersection algorithm selects the most reliable alignment links (high precision), but not all of them (low recall). In contrast to intersection, union results are quite surprising. We have expected recall to be significantly higher than precision, but we have got almost balanced values instead.

As a point of reference we select the Czech-English word alignment as reported by [3]. They evaluated the word alignment quality using precision, recall and alignment error rate (AER) measures, so we do the same for our alignments. Comparing results given by the intersection algorithm, [3] report 27.4% (AER) and we obtain 28.3% (AER) (see Table 2). In case of the union heuristics, results are nearly identical: 25.5% (AER) in [3] and 25.3% (AER) in our experiment. [3] also report the word alignment quality based on lemmatised parallel data. Results of a simmilar experiment for Polish-English language pair are presented in the next section.

Heuristic	Precision	Recall	F_α-measure	AER
intersection	**92.3**	58.6	71.7	28.3
union	71.9	**77.8**	74.7	25.3
grow-diag-final	73.9	76.7	75.3	24.7
grow-diag-final-and (baseline)	77.7	75.8	**76.8**	**23.2**

Table 2. Evaluation of symmetrisation methods of Polish-English bidirectional word alignments.

4.4 Experiment 3: Factor-based Word Alignment

The word alignment training may be based on different features: frequency, part-of-speech, phrase type, the wordform string, etc. Training word alignments on more general word representations such as lemmas or stems seems to perform better for languages with rich morphology. As reported in works describing word alignment experiments carried out with fusional languages (e.g. Czech [3], [4], Romanian [6]), an improvement may be achieved, if we train word alignment on lemmatised parallel data.

Languages used in our experiment represent two language types: Polish is a fusional language with a large number of inflectional word forms; English, in turn, is an isolating language. We suppose that the word alignment quality will improve, if the word alignment is trained on lemmatised word forms.

In order to generate alignments between lemmas, we analyze sentences in the parallel corpus with the available tools and enrich the surface forms with appropriate lemmas. The Polish part of the parallel corpus is analyzed with TaKIPI ([11]), the standard POS-tagger for Polish. Besides POS-tags, TaKIPI provides lemmas the Polish surface forms are annotated with. The English side of the parallel corpus is analyzed using a lemmatiser from the Stanford CoreNLP package[8]. We run Moses on lemma-annotated parallel data. Word alignment links between lemmas are established. We apply the training configuration of IBM models identified previously as the best one: model I (10 iterations), HMM (10 iterations), model III (5 iterations), model IV (5 iterations). Results are given in rows 4-6 of Table 3.

As training on lemmatised parallel data (TaKIPI tagger for Polish, CoreNLP lemmatiser for English) does not yield any remarkable improvement in the quality of the word alignment, we apply another Polish tagger – Pantera [1] to enrich

[8] http://nlp.stanford.edu/sobftware/corenlp.shtml

Factored data	Symmetrisation	Precision	Recall	F_α-measure	AER
Baseline	`gdfa`	77.7	75.8	76.8	23.2
	`intersection`	92.3	58.6	71.7	28.3
	`union`	71.9	77.8	74.7	25.3
PL(TaKIPI) – EN(CoreNLP)	`gdfa`	79.9	76.3	78.0	21.9
	`intersection`	92.1	60.8	73.2	26.7
	`union`	74.7	78.4	76.5	23.5
PL(Pantera) – EN(CoreNLP)	`gdfa`	78.1	74.8	76.4	23.6
	`intersection`	90.3	59.9	72.0	27.9
	`union`	72.7	76.3	74.5	25.5
PL(TaKIPI) – EN(word form)	`gdfa`	78.7	76.0	77.3	22.6
	`intersection`	92.0	61.4	73.7	26.3
	`union`	73.9	77.8	75.8	24.2
PL(Pantera) – EN(word form)	`gdfa`	83.5	80.0	**81.7**	**18.25**
	`intersection`	**96.8**	65.8	78.3	21.6
	`union`	78.3	**81.5**	79.9	20.1

Table 3. Evaluation of the lemma-based word alignment. Explanation: PL – Polish part of the parallel corpus; EN – English side of the parallel corpus; TaKIPI – lemmatising of Polish with TaKIPI tagger; Pantera – lemmatising of Polish with Pantera tagger; CoreNLP – lemmatising of English with Stanford CoreNLP lemmatiser; `gdfa` – `grow-diag-final-and` symmetrisation heuristics.

the Polish tokens with lemmas. We want to verify, if the lemmatising process itself has an impact on results. The English side is annotated with lemmas output by the Stanford CoreNLP lemmatiser again. We repeat the experiment with the same parameter set-up[9]. The only change is the different POS-tagger used to generate Polish lemmas. Results presented in rows 7-9 of Table 3 are even worse than the baseline.

Compared with the Czech-English word alignment based on lemmatised parallel data [3] our results are evidently worse. In case of intersected bidirectional word alignments, [3] report 15.0% (AER) and we get at most 26.7% (AER)[10]. Taking union-symmetrised word alignment into account, [3] reached 17.2% (AER), and we only get 23.5% (AER).

We want to identify the reason for the poor GIZA++ performance on lemmatised parallel data. We compare the *tokens dictionary*[11] against the *lemmas*

[9] Parameter set-up:

- Model training configuration: model I (10 iterations), HMM (10 iterations), model III (5 iterations), model IV (5 iterations);
- Symmetrisation heuristics: `grow-diag-final-and`

[10] We report results of aligning the parallel corpus, with the TaKIPI-lemmatised Polish side and Stanford CoreNLP-lemmatised English side. See 4-6 rows of Table 3.

[11] Tokens dictionary – the list of all surface word forms in a monolingual part of the parallel corpus.

dictionary[12]. We find out that the lemma dictionary provided by TaKIPI lemmatising reduces the number of Polish items the word alignment has been trained on by about 35% (262,705 vs. 169,876) as Table 4 shows. Pantera, in turn, goes even further and reduces training items by about 57% (262,705 vs. 110,828). The lemmatiser offered by Stanford CoreNLP toolkit outputs 142,293 lemmas, which is about 11% more items than in the only tokenised corpus (127,946).

Language	**Tokens**	**Tokens/sentence**	**Tokens dictionary**	**Lemmas dictionary**	
	(total)	(average)	(no of items)	(no of items)	
				TaKIPI	Pantera
Polish	21 560 371	23.9	262 705	169 876	**110 828**
English	24 327 270	26.9	**127 946**	142 293	

Table 4. The Polish-English parallel corpus: size and vocabulary statistics.

As we suppose that the number of items, GIZA++ is trained on, may have a crucial impact on the word alignment quality, we decide to make one more test. The Polish lemmas dictionary generated by Pantera approximates in number to the English tokens dictionary. Therefore we run GIZA++ on parallel data with the tokenised English part and the lemmatised Polish side. Alignment links are generated between English tokens and Polish lemmas. Other parameters are the same as in the previous experiment.

According to our presumptions, unequal number of items in both dictionaries impairs the quality of word alignments trained on this data, i.e., the more approximated size of dictionaries the better word alignment. The word alignment trained on the parallel corpus with the tokenised English side and the Pantera-lemmatised Polish side obtains 81.7% (F_α-measure), which is the best quality we manage to achieve.

5 Conclusion

In order to find out the best parameters for training word alignments on the Polish-English parallel corpus, we carried out series of experiments. The first one identified the best IBM model training configuration (I(10)-HM(10)-III(5)-IV(5)) for generating the Polish-English word alignment. The goal of the second experiment was to find out the symmetrisation method that enables the best combination of bidirectional word alignments. According to our results, `grow-diag-final-and` is the most suitable symmetrisation heuristics for examined languages. In the third experiment we tested the impact of additional factors on word alignment training. We cannot report any crucial improvement while training word alignment on lemmatised data. However, if word alignment

[12] Lemmas dictionary – the list of all single lemmas appeared once or several times in a monolingual part of the parallel corpus.

training was based on the half lemmatised parallel corpus, i.e., the Polish part was Pantera-lemmatised and the English side was tokenised but not lemmatised, the improvement of the word alignment quality was significant. We found out the reason for this, which is the varied size of token/lemma dictionaries. If the number of items in dictionaries for both languages differs significantly, the word alignment quality decreases. There is no doubt that lemmatising the Polish side of the parallel corpus reduces the number of items in the lemma dictionary and approximates the English token dictionary. It remains an open question, if lemmatising the English part of the parallel corpus is necessary. The best Polish-English word alignment achieved 81.7% (F_α-measure) in our experiment.

In a future study, we are going to carry on a detailed error analysis. The identified error sources and problems may help to draft some improvement methods and thus to improve the word alignment quality. Furthermore, we are going to compare the current results with the quality of word alignments trained on a larger parallel corpus. The word-aligned Polish-English parallel corpus will be used to project some linguistic information from English onto Polish.

References

1. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) Advances in Natural Language Processing. Lecture Notes in Computer Science, vol. 6233, pp. 3–14 (2010)
2. Ahrenberg, L., Andersson, M., Merkel, M.: A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. In: Proceedings of the 35 th Annual Meetingof the Association for Computational Linguistics. pp. 29–35 (1998)
3. Bojar, O., Prokopová, M.: Czech-English Word Alignment. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). pp. 1236–1239. ELRA (2006)
4. Bojar, O., Hajič, J.: Phrase-based and deep syntactic English-to-Czech statistical machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation. pp. 143–146. StatMT '08 (2008)
5. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19, 263–311 (1993)
6. Fraser, A., Marcu, D.: ISI's participation in the Romanian-English alignment task. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts. pp. 91–94. ParaText '05 (2005)
7. Fraser, A., Marcu, D.: Measuring Word Alignment Quality for Statistical Machine Translation. Computational Linguistics 33, 293–303 (September 2007)
8. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
9. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of ACL. pp. 177–180. Prague (2007)
10. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
11. Piasecki, M.: Polish Tagger TaKIPI: Rule Based Construction and Optimisation. Task Quarterly 11(1–2), 151–167 (2007)

12. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of LREC 2006. pp. 2142–2147. Genoa, Italy (2006)
13. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing, vol. V, pp. 237–248. Borovets, Bulgaria (2009)
14. Tufiş, D., Barbu, A.M.: Lexical token alignment: Experiments, results and applications. In: Proceedings from The Third International Conference on Language Resources anrd Evaluation (LREC-2002). pp. 458–465. Las Palmas, Spain (2002)
15. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: Proceedings of the 16th conference on Computational linguistics - Volume 2. pp. 836–841. COLING '96 (1996)
16. Yarowsky, D., Ngai, G.: Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In: Proceedings of NAACL 2001. pp. 200–207 (2001)