

Automatically Inducing a Part-of-Speech Tagger by Projecting from Multiple Source Languages Across Aligned Corpora

Victoria Fossum¹ and Steven Abney²

¹ Dept. of EECS, University of Michigan, Ann Arbor MI 48105
vfossum@umich.edu

² Dept. of Linguistics, University of Michigan, Ann Arbor MI 48105
abney@umich.edu

Abstract. We implement a variant of the algorithm described by Yarowsky and Ngai in [21] to induce an HMM POS tagger for an arbitrary target language using only an existing POS tagger for a source language and an unannotated parallel corpus between the source and target languages. We extend this work by projecting from *multiple* source languages onto a single target language. We hypothesize that systematic transfer errors from differing source languages will cancel out, improving the quality of bootstrapped resources in the target language. Our experiments confirm the hypothesis. Each experiment compares three cases: (a) source data comes from a single language A, (b) source data comes from a single language B, and (c) source data comes from both A and B, but half as much from each. Apart from the source language, other conditions are held constant in all three cases – including the total amount of source data used. The null hypothesis is that performance in the mixed case would be an average of performance in the single-language cases, but in fact, mixed-case performance always exceeds the maximum of the single-language cases. We observed this effect in all six experiments we ran, involving three different source-language pairs and two different target languages.

1 Introduction

1.1 Background

Statistical NLP techniques typically require large amounts of annotated data. Labelling data by hand is time-consuming; a natural goal is therefore to generate text analysis tools automatically, using minimal resources. Yarowsky et al. [22] present methods for automatically inducing various monolingual text analysis tools for an arbitrary target language, using only the corresponding text analysis tool for a source language and a parallel corpus between the source and target languages. Hwa et al. [15] induce a parser for Chinese text via projection from English using a similar method to that of [22]. Cucerzan and Yarowsky [8] present a method for bootstrapping a POS tagger for an arbitrary target language using

only a bilingual dictionary between the source and target languages, a “basic library reference grammar” for the target language, and an existing corpus in the target language.

While automatically induced text analysis tools use fewer resources, their accuracy lags behind that of more resource-intensive tools. One solution to the problem of error reduction on NLP tasks is to train multiple classifiers, then compute a consensus classifier. Combining multiple classifiers is an effective way to reduce error if the errors made by each classifier are independently distributed. Such approaches have been successfully applied to a range of NLP tasks. Brill and Wu [4], van Halteren et al. [19], and Zavrel and Daelemans [23] investigate various methods for improving the performance of statistical POS taggers by combining multiple such taggers. Henderson and Brill [14] combine the Charniak, Collins, and Ratnaparkhi parsers to achieve an accuracy surpassing the best previous results on the WSJ. Gollins and Sanderson [10] apply projection via multiple source languages to reduce error in cross-linguistic information retrieval.

1.2 Motivation

We hypothesize that a large component of the error rate in the automatically induced text analysis tools generated by [22] is due to morphosyntactic differences between the source and target languages that are specific to each source-target language pair. Therefore, training POS taggers on additional source languages should result in multiple classifiers which produce independently distributed errors on the target language.

Previous research in classifier combination for POS tagging has focused primarily on combining various statistical classifiers trained on data in the same language. Thus, our approach is novel in its exploitation of differences across languages, rather than differences across statistical methods, to improve performance on POS tagging. Our method is general in that it does not rely on language-specific information, and requires no annotated resources in the target language.

Our method is easily extensible to new languages. While it requires a parallel corpus between each source language and the target language, the corpora used to train each single-source tagger need not be translations of the same text. Furthermore, our algorithm is applicable even to target languages belonging to distinct language families from those of the source languages.

1.3 Task Overview

Using existing POS taggers for English, German, and Spanish, we generate single-source taggers for Czech and French via projection across parallel translations of the Bible. To obtain a theoretical upper bound on the performance improvement that is possible by combining multiple POS taggers, we measure the *complementarity* between each pair of single-source taggers. We examine

various ways to combine the output of these single-source taggers into a consensus tagger, and measure the resulting performance improvement.

2 Methods

2.1 Single-Source POS Tagger Induction

We implement a variant of the algorithm described in [21] for constructing a single-source bigram-based HMM POS tagger for a target language. First, we identify a language (the “source language”) for which a POS tagger exists, and a sentence-aligned parallel corpus consisting of text in the source language and its translation in the target language. We then align the parallel corpora at the word-level using GIZA++ [1]. Next, we annotate the source text using an existing POS tagger. Finally, we project these annotations across the parallel text from the source text to the target text, smooth these projections, and use the projected annotations to train an HMM POS tagger for the target language.

In more detail, we implement the following procedure, based on [21]:

1. Obtain a sentence-aligned parallel corpus in the source and target languages (see Section 2.4).
2. Align the parallel corpus at the word-level using GIZA++¹.

English: He(1) likes(2) cats(4).

French: Il(1) aime(2) les(3) chats(4).

3. Tag the source portion of the parallel corpus using an existing POS tagger for the source language². We use the Brill tagger³ for English [3], the TNT tagger⁴ for German, and the SVMTool tagger⁵ for Spanish.

English: He/PRP likes/VBP cats/NNS.

Since the POS tagger for each source language uses its own distinct tagset, we convert the output of each tagger to a “generic” tagset for comparison purposes. Additionally, we label each POS tag as belonging to one of several more general “core” tagset categories (see Table 1).

4. Using the mapping induced by the word-level alignments, project the POS tags from the source language onto the target language.

French: Il/PRP aime/VB les/NULL chats/NNS.

¹ GIZA++ is a component of EGYPT, an open-source implementation of IBM’s statistical machine translation system [1].

² For all existing POS taggers, we use the default models provided with the tagger for training in the source language. For taggers with variable parameter settings, we use the default settings for all parameters.

³ A transformation-based tagger [3].

⁴ A bigram-based Markov tagger[2].

⁵ An SVM-based tagger [9].

Note that tag projection is complicated by the occurrence of many-to-one word alignments from source to target. To handle such cases, we compute two estimates of tag probabilities, $P(t_i|w_i)$: one using only 1-to-1 alignments, and the other using 1-to-n alignments. We then linearly combine the two estimators.

5. Before computing the $P(w_i|t_i)$ model, several steps must be taken to smooth the initial, noisy tag projections. First, $P(w_i|t_i)$ can be decomposed as follows:

$$P(w_i|t_i) = \frac{P(t_i|w_i) * P(w_i)}{P(t_i)}$$

To smooth $P(t_i|w_i)$, the simplifying assumption is made that in most natural languages, each word has at most two possible POS tags at the core tagset granularity. We count the relative frequency of each tag that is assigned to that French word by the tag projection from English, then discard all but the two most frequently assigned core tags. We then recursively smooth the tag probabilities in favor of the two most probable subtags for each of the core tags, where the subtags are members of the more finely grained “generic” tagset. We compute $P(t_i)$ and $P(w_i)$ using corpus frequency.

6. We estimate the probability of unknown words using the probability of words appearing only once in the training corpus. We replace all words occurring only once in the training corpus by the “UNK” token.

7. Before computing the $P(t_j|t_i)$ model, we filter the training data to remove those sentence pairs whose alignment score (as determined by GIZA++) falls into the lowest 25% of alignment scores. To estimate the probability of unknown state transitions, we perform Witten-Bell smoothing [20] on $P(t_j|t_i)$ to assign non-zero probabilities to state transitions not seen in the training data.

8. The resulting model defines an HMM bigram-based tagger in the target language. We use the Viterbi algorithm to determine the most likely sequence of tags given a sentence in the target language [17].

2.2 Multiple-Source POS Tagger Induction

To compute a multiple-source consensus tagger, we train n single-source taggers using n parallel texts, each pairing one of the source languages with the target language. We then apply each single-source tagger to the test sentences. For each word in the test sentences, we record the probability distribution $P_i(t|w)$ over possible tags that the i^{th} single-source tagger produces. We then compute two consensus taggers, *Majority Tag* and *Linear Combination*, by combining the output from each of the n taggers, $P_1(t|w) \dots P_n(t|w)$ as follows:

Majority Tag: Each tagger outputs the most likely tag

$$t_i^{\text{best}} = \underset{t}{\operatorname{argmax}}(P_i(t|w))$$

for w . We select the tag from $t_1^{\text{best}}, \dots, t_n^{\text{best}}$ that receives the greatest number of votes from single-source taggers. To break ties, we select the tag chosen with the highest probability by the taggers that selected it.

Linear Combination: Each tagger outputs a vector of probabilities over possible tags t given w . We take a linear combination of these vectors to compute $P_{linear}(T|w)$, then select the tag t_{linear} with the highest probability.

$$P_{linear}(T|w) = \sum_{i=1}^n k_i * (P_i(T|w))$$

$$t_{linear} = \operatorname{argmax}_t (P_{linear}(t|w))$$

In our experiments, we set $k_i = \frac{1}{n}$, so we effectively average the probability distributions of each tagger over possible tags t for w .

2.3 Tagsets

Two tagsets of different granularities are used in the experiments: the coarse-grained “core” and fine-grained “generic” tagsets (see Table 1). While it can be difficult to map fine-grained POS tags from one language directly onto another because of morphological differences between languages, languages tend to agree on tags at a coarse-grained level.

2.4 Data Sets

We use two corpora in our experiments: the Bible (with translations in English, Czech, French, German, and Spanish), and the Hansards parallel corpus of Canadian parliamentary proceedings (with translations in English and French). For

Table 1. Generic and Core Tagsets

POS	Generic	Core
Noun	NN	N
Proper Noun	NNP	N
Verb, Inf.	VB	V
Verb, Present	VBP	V
Verb, Present Part.	VBG	V
Verb, Past Part.	VBN	V
Verb, Past	VBD	V
Determiner	DT	D
Wh-Determiner	WDT	D
Conjunction	CC	C
Number	CD	NUM
Adverb	RB	R
Wh-Adverb	WRB	R
Adjective	JJ	J
Pronoun	PRP	P
Preposition	IN	I

the Bible experiments, we use the entire 31,100-line text: training data consists of either one 31,000-line excerpt or two 15,500-line excerpts, while testing data consists of a held-out 100-sentence excerpt. For the Hansards experiments, training data consists of a 85,000-line excerpt; testing data consists of a held-out 100-sentence excerpt.

We perform the following pre-processing steps. Each text is filtered to remove punctuation and converted to lower case; accents are preserved. The English, French, German, and Spanish texts are tokenized to expand elisions.⁶

3 Results

We report percent agreement with the correct tags, determined by comparison with the output of the Treetag tagger⁷ for French, and a hybrid rule-based/HMM-based tagger⁸ for Czech. For French, agreement with the correctly tagged text is measured on the generic and core tagsets. For Czech, agreement is measured on the core tagset only, since this is the POS tagset provided by the tagger we use for evaluation purposes. All experiments use 5-fold cross-validation. For each iteration, the parallel corpus is divided randomly into training and testing sets. The accuracy of each single-source tagger is limited by the accuracy of the tagger used to tag the source training text; the accuracy of the evaluation of each tagger’s performance on French and Czech text is limited by the accuracy of the reference tagger against which it is compared (Table 2).

Table 2. Reported Accuracy of Existing POS Taggers used to Train Single-Source Taggers

Tagger	Language	% Accuracy	F-measure	Test Corpus
Brill	English	96.6%	—	Penn Treebank (English)
TNT	German	—	—	—
SVMTool	Spanish	96.89%	—	LEXESP (Spanish)
TreeTag	French	96.36%	—	Penn Treebank (English)
Rules + HMM	Czech	—	95.38%	PDT (Czech)

3.1 Single-Source

For each single-source tagger, we train on 31,000 lines of the parallel Bible between the source and target languages and test on 100 held-out lines of the Bible in the target language. We report the accuracy of the induced taggers on French (Tables 5 and 4) and Czech (Table 6).

⁶ e.g. “doesn’t” → “does not”, “qu’il” → “que il”, “zum” → “zu dem”, and “del” → “de el”. This tokenization represents the only step of our algorithm that requires additional language-specific knowledge beyond the resources already given.

⁷ A decision-tree-based tagger [18].

⁸ [13].

To compare our baseline single-source tagger performance against that of [21], we conduct the following experiment, after the experimental procedure used by [21]. We train a single-source English-projected tagger for French on a 2,000,000-word (approximately 85,000-line) excerpt of the French-English Hansards corpus and test it on a 100-line excerpt of the same corpus. We obtain accuracies of 86.5% and 91.1% on the generic and core tagsets, respectively; [21] report accuracies of 91% and 94% on the “English Equivalent” and “core” tagsets, respectively.⁹

3.2 Multiple-Source

Complementarity: We compute the pairwise *complementarity* of each pair of single-source taggers. Brill and Wu [4] define the complementarity of a pair of taggers i and j as the percentage of cases when tagger i is wrong that tagger j is correct (See Table 3):

$$\text{Comp}(i, j) = \left(1 - \frac{\text{errors}_i \cup \text{errors}_j}{\text{errors}_i}\right) * 100$$

Table 3. Complementarity (row,col) of Single-Source Taggers

Source	French Bible						Czech Bible		
	Generic Tagset			Core Tagset			Core Tagset		
	English	German	Spanish	English	German	Spanish	English	German	Spanish
English	0	38.95	32.87	0	32.75	37.13	0	22.08	18.71
German	42.40	0	44.93	30.49	0	38.95	15.47	0	17.31
Spanish	41.12	48.83	0	35.64	39.51	0	19.95	24.98	0

Pairwise Combination: To determine whether tagger performance improves by using training data from two different source languages, without increasing the total amount of training data, we perform the following experiments. For each possible combination of two single-source taggers, we partition the Bible into two 15,500-line training sets (the first, a parallel corpus between one source language and the target language; the second, a parallel corpus between the other source language and the target language), and a 100-line held-out testing set. We train the first single-source tagger on one half, train the second single-source tagger on the second half, combine their output using the methods described in Section 2.2, and test the resulting consensus tagger on a held-out 100-line excerpt of the French (Tables 4 and 5) or Czech (Table 6) Bibles. For each pairwise combination of taggers, we report the percent error reduction of the combined tagger in comparison to the average accuracy of the constituent single-source taggers.

⁹ Our “generic” and “core” tagsets correspond approximately to the “English Equivalent” and “core” tagsets used by [21]. Since we do not have access to the same testing set used by [21], we report results on a held-out excerpt of the Hansards corpus.

Table 4. % Accuracy of Single-Source, Pairwise-Combined, and n -way Combined Taggers Using Generic Tagset on French Bible

Sources	% Accuracy		% Error Rate Reduction
	Linear	Majority	
English	81.95	81.95	–
German	81.21	81.21	–
Spanish	79.76	79.76	–
Eng. + Ger.	84.52	84.30	15.96
Eng. + Span.	84.42	84.48	18.91
Ger. + Span.	83.89	84.09	18.45
E. + G. + S.	85.80	85.61	25.38

Table 5. % Accuracy of Single-Source, Pairwise-Combined, and n -way Combined Taggers Using Core Tagset on French Bible

Sources	% Accuracy		% Error Rate Reduction
	Linear	Majority	
English	85.67	85.67	–
German	86.66	86.66	–
Spanish	86.54	86.54	–
Eng. + Ger.	88.06	88.05	13.67
Eng. + Span.	88.13	88.12	14.54
Ger. + Span.	89.12	89.19	19.33
E. + G. + S.	89.87	89.43	26.11

n -Way Combination: To examine how much tagger performance can be improved by increasing the total amount of training data n -fold and training each of n single-source taggers on the full 31,000 lines of the Bible, then computing a consensus tagger, we perform the following experiment. We train each single-source tagger on 31,000 lines of the Bible, then compute the consensus output of all 3 single-source taggers on a held-out 100-line excerpt of the French (Tables 4 and 5) or Czech (Table 6) Bibles. For each n -way combination of taggers, we report the percent error reduction of the combined tagger in comparison to the average accuracy of the constituent single-source taggers.

4 Discussion

All multiple-source taggers outperform the corresponding single-source taggers—thus, incorporating multiple source languages improves performance, even when the total amount of training data is held constant (as in the pairwise combination experiments).

4.1 Single-Source Taggers

We expect performance to be highest for those source-target language pairs that are most similar to each other, linguistically. At the generic tagset level, the

Table 6. % Accuracy of Single-Source, Pairwise-Combined, and n -way Combined Taggers Using Core Tagset on Czech Bible

Sources	% Accuracy		% Error Rate Reduction
	Linear	Majority	
English	62.53	62.53	–
German	65.27	65.27	–
Spanish	63.27	63.27	–
Eng. + Ger.	65.44	65.98	5.76
Eng. + Span.	65.41	65.28	6.77
Ger. + Span.	67.18	67.75	9.74
E. + G. + S.	67.13	67.36	10.12

poor performance of the Spanish-projected single-source tagger on French text is partially due to a discrepancy between the SVMTool tagset [9] and our generic tagset¹⁰. At the core tagset level, the distinction between verb tenses becomes irrelevant, and the performance of the Spanish-projected tagger matches that of the other single-source taggers more closely on French data; still, its performance is lower than expected given the close morphosyntactic correspondence between Spanish and French.¹¹

For several reasons, we expect single-source tagger performance to be poorer on Czech (Table 6) than on French (Tables 5 and 4). First, Czech is a “highly inflected” language: the role of function words in the Germanic and Romance languages is typically filled by suffixes in Czech. Second, Czech exhibits a “relatively free word order” [7]. Since a great deal of the POS information exploited by an HMM tagger is contained in sequences of function words¹², these features of Czech hinder the performance of an HMM POS tagger.¹³ Finally, Czech belongs to the Slavic language family, and is therefore further removed than French from the Germanic and Romance families of the source languages used to train the single-source taggers.

Although our single-source taggers do not replicate the performance results reported by [21] (91% and 94% accuracy on generic and core tagsets, respectively), our primary concern is not their absolute performance but rather their

¹⁰ e.g., SVMTool [9] does not make certain distinctions in verb tense that we make in our generic tagset.

¹¹ One likely explanation for this discrepancy is that we do not optimize the parameters of the Spanish POS tagger used to annotate the source corpus to suit the input format of our data set, but instead use the default settings. We estimate that optimizing these parameters to match our data set could result in an increase of 1-2% accuracy in the Spanish-projected source tagger for French and Czech; however, such an increase in performance of one of the baseline experiments would not change our conclusion in a significant way.

¹² e.g., a “DT” is likely to be followed by a “NN” in English.

¹³ The Czech tagger we use for reference [13] combines a rule-based morphological analyzer with an HMM POS tagger to combat these problems; our induced HMM POS taggers, lacking any morphological analysis component, may not exploit the correct type of information for such languages.

performance relative to the multiple-source taggers. We think it plausible that the improvements we observe would also be observed with Yarowsky’s single-source taggers, but it remains an open question.

4.2 Multiple-Source Taggers

Complementarity. Pairwise complementarity among single-source taggers is relatively high on French at both tagset granularities (Table 3). The low pairwise complementarity of taggers on Czech may indicate the existence of a ceiling on the performance of the single-source tagger induction algorithm, imposed by the limited degree of similarity between any of the source languages with the target language. Even under such circumstances, we still see improvement (though diminished) by combining single-source taggers for Czech.

One factor whose influence upon tagger complementarity must be acknowledged is the diversity of the statistical models underlying each of the POS taggers used to tag the source portion of the training text. Since we use a different type of tagger to tag each source language, we cannot separate the component of complementarity that is caused by the difference in statistical models among sources from the component caused by the difference in languages.

Pairwise Combination. All pairwise combined taggers outperform the corresponding single-source taggers, though the total amount of training data is unchanged. We observe this improvement on both French and Czech. This suggests that our approach is likely to improve performance over single-source taggers on a wide range of target languages, and does *not* depend upon a close correspondence between any of the source and target languages.

n -Way Combination. As expected (given the n -fold increase in training data), all n -way combined taggers outperform the corresponding single-source taggers, suggesting that when parallel training data between a particular source-language pair is limited, the performance of a POS tagger projected across that language pair can be improved by the use of a parallel corpus between the target language and a different source language.

5 Conclusion

Projection from multiple source languages significantly improves the performance of automatically induced POS taggers on a target language. We observe performance gains from incorporating multiple source taggers even when the total amount of training data is held constant, indicating that multiple languages provide sources of information whose errors are independent and randomly distributed to a large extent. The approach presented here is general in that it does not depend on any language-specific resources in the target language beyond parallel corpora. Our results suggest that the performance of text analysis tools induced using parallel corpora can benefit from the incorporation of resources in other languages, even in the case of source languages belonging to distinct linguistic families from the target language.

6 Future Work

To further improve the accuracy of induced multiple-source taggers, we plan to investigate other methods for combining the output of single-source POS taggers. We hypothesize that combining the models constructed by each tagger *before* applying each tagger to the testing set would result in greater performance gains.

References

1. Yasser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah Smith and David Yarowsky: Statistical machine translation. Johns Hopkins University 1999 Summer Workshop on Language Engineering (1999)
2. Thorsten Brants: TnT – a statistical part-of-speech tagger. In Proceedings of the 6th Applied NLP Conference, ANLP-2000, April 29 – May 3, 2000, Seattle, WA. (2000)
3. Eric Brill: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics Vol. 21 No. 4 (1995) 543-565
4. Eric Brill and Jun Wu: Classifier Combination for Improving Lexical Disambiguation. Proceedings of the ACL (1998)
5. Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin: A Statistical Approach to Machine Translation. Computational Linguistics Vol. 16 No. 2 (1990) 79–85
6. S. Clark, J. Curran, and M. Osborne: Bootstrapping POS taggers using unlabelled data. In Walter Daelemans and Miles Osborne, editors, Proceedings of CoNLL-2003, Edmonton, Canada (2003) 49–55
7. Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann: A Statistical Parser for Czech. Proceedings of the 37th Annual Meeting of the ACL, College Park, Maryland (1999)
8. Silviu Cucerzan and David Yarowsky: Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day. Proceedings of the Sixth Conference on Natural Language Learning (CoNLL) (2002)
9. Jesus Gimenez and Lluís Marquez: SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal (2004)
10. Tim Gollins and Mark Sanderson: Improving Cross Language Information Retrieval with Triangulated Translation. Proceedings of the 24th annual international ACM SIGIR conference 90–95 (2001)
11. French-English Hansards Corpus of Canadian Parliamentary Proceedings.
12. Jan Hajic and Barbora Hladka: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset, COLING-ACL (1998) 483–490
13. Jan Hajic, Pavel Krbeč, Pavel Kevton, Karel Oliva, and Vladimír Petkevič: Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. Proceedings of the ACL (2001)
14. John C. Henderson and Eric Brill: Exploiting Diversity in Natural Language Processing: Combining Parsers. Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (1999) 187–194

15. Rebecca Hwa, Philip Resnik, and Amy Weinberg: Breaking the Resource Bottleneck for Multilingual Parsing. Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data (2002)
16. Gideon Mann and David Yarowsky: Multipath translation lexicon induction via bridge languages. In Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (2001) 151–158
17. Lawrence Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE Vol. 77 No. 2 (1989)
18. Helmut Schmid: Probabilistic Part-of-Speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing, Manchester, UK. (1994)
19. Hans van Halteren, Jakub Zavrel, and Walter Daelemans: Improving Data Driven Wordclass Tagging by System Combination. Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics (1998) 491–497
20. Ian Witten and Timothy Bell: The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions in Information Theory, Vol. 37 No. 4 1085–1094 (1991)
21. David Yarowsky and Grace Ngai: Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. Proceedings of NAACL (2001) 200–207
22. David Yarowsky, Grace Ngai, and Richard Wicentowski: Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. Proceedings of HLT (2001)
23. Jakub Zavrel and Walter Daelemans: Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers. Proceedings of LREC-2000, Athens (2000)