Morphological Tagging: Data vs. Dictionaries

Jan Hajič*
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
hajic@cs.jhu.edu

Abstract

POS tagging for English seems to have reached the the human levels of error, but full morphological tagging for inflectionally rich languages, such as Romanian, Czech, or Hungarian, is still an open problem, and the results are far from being satisfactory. This paper presents results obtained by using an universalized exponential feature-based model for five such languages. It focuses on the data sparseness issue, which is especially severe for such languages (the more so that there are no extensive annotated data for those languages). In conclusion, we argue strongly that the use of an independent morphological dictionary is the preferred choice to more annotated data under such circumstances.

1 Full Morphological Tagging

English Part of Speech (POS) tagging has been widely described in the recent past, starting with the (Church, 1988) paper, followed by numerous others using various methods: neural networks (Julian Benello and Anderson, 1989), HMM tagging (Merialdo, 1992), decision trees (Schmid, 1994), transformation-based error-driven learning (Brill, 1995), and maximum entropy (Ratnaparkhi, 1996), to select just a few. However different the methods were, English dominated in these tests.

Unfortunately, English is a morphologically "impoverished" language: there are no complicated agreement relations, word order variation is minimal, and the morphological categories are either extremely simple (`-s` for plural of nouns, for example), or (almost) nonexistent (cases expressed by inflection, for example) - with not too much exceptions and irregularities. Therefore the number of tags selected for a English tagset is not that large (40-75 in the typical case). Also, the average ambiguity is low (2.32 tags per token on the manually tagged Wall Street Journal part in the Penn Treebank, for example).

Highly inflective and agglutinative languages are different. Obviously one can limit the number of tags to the major part-of-speech classes, plus some (like the Xerox Language Tools (Chanod, 1997) for such languages do), and in fact achieve similar performance, but that limits the usefulness of the results so obtained for further analysis. These languages, obviously, do not use the rich inflection just for the amusement (or, embarrassment) of their speakers (or, NLP researchers): the inflectional categories carry important information which ought to be known at a later time (e.g., during parsing). Thus one wants not only to tell apart verbs from nouns, but also nominative from genitive, masculine animate from inanimate, singular from plural – all of them being often ambiguous one way or the other.

The average tagset, as found even in a moderate corpus, contains between 500 and 1,000 distinct tags – whereas the size of the set of possible and plausible tags can reach 3,000 to 5,000. Obviously, any of the statistical methods used for English (even if fully supervised) clash with (or, fall through) the data sparseness problem (see below Table 1 for details).

There have been attempts to solve this problem for some of the highly inflectional European languages ((Daelemans et al., 1996), (Erjavec et al., 1999), (Tufis, 1999), and also our own in (Hajič and Hladká, 1997), (Hajič and Hladká, 1998), see also below), but so far no method nor a tagger has been evaluated against a larger number of those languages in a similar setting, to allow for a side-by-side comparison of the difficulty (or ease) of full morphological tagging of those languages. Thanks to the MULTEXT-EAST project (Véronis, 1996a), there are now five annotated corpora available (which are manually, fully morphologically tagged) to perform such experiments.

2 The Languages Used and The Training Data

We use the Multext-East-annotated version of the Orwell's 1984 novel in Czech, Estonian, Hungarian,

* The work described herein has been started and largely done within author's home institution, the Institute of Formal and Applied Linguistics, Charles University, Prague, CZ, within the project VS96151 of the Ministry of Education of the Czech Republic and partially also under the grant 405/96/K214 of the Grant Agency of the Czech Republic.

Romanian and Slovene.[1]. The annotation uses a single SGML-based formal scheme, and even common guidelines for tagset design and annotation, but the tagsets differ nevertheless substantially as the languages differ as well: Romanian is a French-like romance language, Hungarian is agglutinative, and the other languages are more or less inflectional-type languages[2]. The annotated data contains about 100k tokens (including punctuation) for each language; out of those, the first 20k tokens has been used for testing, the rest for training. We have also extended the tag identifiers by appending a string of hyphens ('-') to suit the exponential tagger which expects the tags to be of equal length; the mapping was 1:1 for all tags in all languages, since the "long" tags are in fact the Multext-East standard.

From the tagging point of view, the language characteristics displayed in Table 1 are the most relevant[3].

3 The Methodology

The main tagger used for the comparison experiment is the probabilistic exponential-model-based, error-driven learner we described in detail in (Hajič and Hladká, 1998). Modifications had to be made, however, to make it more universal across languages.

3.1 Structure of the Model

The model described in (Hajič and Hladká, 1998) is a general exponential (specifically, a log-linear) model (such as the one used for Maximum Entropy-based models):

$$p_{AC}(y|x) = \frac{\exp(\sum_{i=1}^{n} \lambda_i f_i(y,x))}{Z(x)} \qquad (1)$$

where $f_i(y,x)$ is a binary-valued *feature* of the event value being predicted and its context, λ_i is a weight of the feature f_i, and the $Z(x)$ is the natural normalization factor. This model is then essentially reduced to Naive Bayes by the approximation of the λ_i parameters, which is done because there are millions of possible features in the pool and thus the full entropy maximization is prohibitively expensive, if one wants to select a small number of features instead of keep them all.

The tags are predicted separately for each morphological category (such as POS, NUMBER, CASE, DEGREE OF COMPARISON, etc.). The model makes an extensive use of so-called "ambiguity classes" (ACs). An ambiguity class is a set of values (such as genitive and accusative) of a single category (such as CASE) which arises for some word forms as a result of morphological analysis. For unambiguous word forms (unambiguous from the point of view of certain category), the ambiguity class set contains only a single value; for ambiguous forms, there are 2 or more values in the AC. For example, let's suppose we use part-of-speech (POS), number and tense as morphological categories for English; then the word form "`borrowed`" is 2-way ambiguous in POS ({`V,N`} for verb and noun, respectively), unambiguous in number (linguistic arguments apart, number is typically regarded "not applicable" to adjectives as well as to almost all forms of verbs in English), and 3-way ambiguous in tense ({`P,N,-`} for past tense, past participle, and "not applicable" in the adjective form).

The predictions of the models are always conditioned on the ambiguity class of the category (POS, NUMBER, ...) in question. In other words, there is a separate model for each category *and* an ambiguity class from that category. Naturally, there is no model for unambiguous ACs classes. However, even though the ambiguity classes bring very valuable information about the word form being tagged and a reliable information about the context (since they are fixed during tagging), using ACs causes also an unwelcomed effect of partitioning the already scarce data and also effectively ignores statistics of the unambiguous cases.

The context of features uses the neighboring words (original word forms) and ambiguity classes on subtags, where their relative position in text might be either fixed (0, -1, +1) or "variable" using a value of the POS subtag as the "stop here" criterion, up to 4 text positions (words) apart.

3.2 General Subtag Features

The original model uses the ambiguity classes not only for conditioning on context in features, but also for the individual models based on category *and* an AC.

More general features have been introduced, which do not depend on the ambiguity class of the subtag being predicted any more. This allows to learn also from unambiguous tokens. However, the training time is increased dramatically by doing so since *all* events in the training data have to be taken into consideration, as opposed to the case of training

[1]There are more languages involved in the Multext-East project, but only these five languages have been really carefully tagged; English is unfortunately tagged using Eric Brill's tagger trained in unsupervised mode, leaving multiple output at almost every ambiguous token, and Bulgarian is totally unusable since it has been tagged automatically with only a baseline tagger. The English results reported below thus come from the Penn Treebank data, from which we have used roughly 100,000 words to match the training data sizes for the remaining languages. For Czech, Hungarian, and Slovene we use later versions of the annotated data (than those found on the Multext-East CD) which we obtained directly from the authors of the annotations after the Multext-CD had been published, since the new data contain rather substantial improvements over the originally published data.

[2]For detailed account of the lexical characteristics of these languages, see (Véronis, 1996b).

[3]We have included English here for comparison purposes, since these characteristics are independent of the annotation.

Table 1: Training data in numbers

Language	Training Size	Tagset Size	Ambiguous Tokens
English[4]	99903	139	38.65%
Czech	87071	970	45.97%
Estonian	81383	476	40.24%
Hungarian	102992	401	21.58%
Romanian	104583	486	40.00%
Slovene	94457	1033	38.01%

the small AC-based model, when only those training events which contain the particular AC are used.

3.3 Variable Distance Condition

The "stop" criterion for finding the appropriate relative position was originally based on hard coded choices suitable for the Czech language only, and of course it depended on the tagset as well. This dependency has been removed by selecting the appropriate conditions automatically when building the pool of possible features at the initialization phase[5] (using the relative frequency of the POS ambiguity classes, and a threshold to cut off less frequent categories to limit the size of the feature pool).

3.4 Weight Variation

Even though the full computation of the appropriate feature weight is still prohibitive (the more so when the general features are added), the learner is now allowed to vary the weights (in several discrete steps) during feature selection, as a (somewhat crude) attempt to depart from the Naive Bayes simplification to the approximation of the "correct" Maximum Entropy estimation.

3.5 Handling Unknown Words

In order to compare the effects of (not) using an independent dictionary, we have added an unknown word handling module to the code.[6] It extracts the prefix and suffix frequency information (and the combination thereof) from the training data. Then, for each of the combinations selects the most frequent set of tags seen in the training data and stores it for later use. When tagging, the data is first piped through a "guesser" which assigns to the unknown words such a set of possible tags which is stored with the longest matching prefix/suffix combination.

[5]Also, the use of variable-distance context may be switched off entirely.

[6]Originally, the code relied exclusively on the use of such an independent dictionary. Since the coverage of the Czech dictionary we have used is extensive, we have been simply ignoring the unknown word problem altogether in the past.

4 The Results

4.1 Reporting Error Rate: Words vs. Tokens

Since "best-only" tagging has been carried out, the error rate (i.e, 100 - accuracy in %) measure has been used throughout as the only evaluation criterion. However, since some results reported previously apparently use only the "real" words as the basis for accuracy evaluation, whereas other count every token (including punctuation[7], for example), we have computed both and report them separately[8].

4.2 Availability of Dictionary Information

We use two methods of obtaining the set of possible tags for (i.e., for morphological analysis of) any given word form (which includes the handling of unknown words). First, we use only information which may be obtained automatically from the manually annotated corpus (we call this method *automatic*). This is the way the Maximum Entropy tagger (Ratnaparkhi, 1996) runs if one uses the binary version from the website (see the comparison in the Section 5).

However, it is not unreasonable to assume that a larger *independent dictionary* exists, which can help to obtain a list of possible tags for each word form in test data. This is what we have at our disposal for the languages in question, since the development of such a dictionary was part of the Multext-East project. We can thus assume a dictionary info is available for unknown words in the test data, i.e., even though there is no statistics available for them (since they did not appear in the training data), all possible tags for (almost[9]) every test token *are* available. This method is referred to as *independent* in the following text.

We have also used a third method of obtaining a dictionary information (called *mixed*), namely, by using only the words from the training data, but

[7]And sometimes a separate token for sentence boundary

[8]The Table 1 has been computed using all tokens. In fact, the languages differ significantly in the proportion of punctuation: from about 18% (English) to 30% (Estonian).

[9]Depending on the quality of the independent dictionary. Of course, the tagsets must match, which could be a problem per se. Here it is simple, since the dictionaries have been developed using the same tagsets as the tagged data.

"completing" the information obtained about them from the training data by including **all** other possible tags for such words. Therefore the net result is that during testing, we have only training words at our disposal, but with a complete dictionary information (as if coming from a full morphological dictionary)[10].

The results on the full training data set are summarized in Table 2.

The baseline error rate is computed as follows. First of all, we use the independent dictionary for obtaining the possible tags for each word. Then we extract only the lexical information from the current position[11] and counts used for smoothing (which is based on the ambiguity classes only and it does not use lexical information). The system is then trained normally, which means it uses the lexical information only if the AC-based smoothing[12] alone does not work. This baseline method is thus very close to the usual baseline method of using simple conditional distribution of tags given words.

1.91.0

The message of Table 2 seems to be obvious; but before we jump to conclusions, let's present another set of experiments.

In view of recent interest in dealing with "small languages", and with regard to the questions of cost-effectiveness of using the "human" resources (i.e. annotation vs. rule-writing vs. tools development etc.), we have also performed experiments with reduced training data size (but with an enriched feature pool – by lowering thresholds, adding more of "general features" as described above, etc. – as allowed by reasonable time/space constraints).[13]

These results are summarized in Table 3 (using only dictionary derived from the training data), Table 4 (using words from training data with morphological information completed from a dictionary) and Table 5 (using the "independent" dictionary). In all cases, we again count only true words (no punctuation). Accordingly, the major Part-of-speech error rate is reported, too (12 tags to be distinguished only: Noun, Verb, Adjective, ...; see Tables 6, 7, and 8).

[10]This arrangement removes the "closed vocabulary" phenomenon from the test data, since for the Multext-East data, we did not have a truly independent vocabulary available.

[11]Words from the training data which are not singletons (freq > 1) are used. Surprisingly enough, it would not hurt to use them too. We believe it is due to the smoothing method used. Even though this is valid only for the baseline experiment, we have observed in general that this form of exponential model (with error-driven training, that is) is remarkably resistant to overtraining.

[12]Using ACs linearly interpolated with global unigram subtag distribution and finally the uniform distribution.

[13]By reasonable we mean less than a day of CPU for training.

Table 9: Exponential w/feature selection vs. Maximum Entropy tagger (Words-only Error Rate, no dictionary)

Language	Tagger	
	Exp.	MaxEnt
English	9.18%	6.38%
Czech	18.83%	17.77%
Estonian	13.95%	14.92%
Hungarian	8.16%	8.55%
Romanian	7.76%	7.66%
Slovene	16.26%	17.44%

4.3 Tagger Comparison

The work (Erjavec et al., 1999) consistently compares several taggers (HMM, Brill's Transformation-based Tagger, Ratnaparkhi's Maximum Entropy tagger, and the Daelemans et al.'s Memory-based Tagger) on Slovene. We have chosen the Maximum Entropy tagger (Ratnaparkhi, 1996) for a comparison with our universal tagger, since it achieved (by a small margin) the best overall result on Slovene as reported there (86.360% on all tokens) of taggers available to us (MBT, the best overall, was not freely available to us at the time of writing). We have trained and tested the Maximum Entropy Tagger on exactly the same data, using the off-the-shelf (java binary only) version.

The results are compared in the Table 9.

Since we want to show how a tagger accuracy is influenced by the amount of training data available, we have run a series of experiments comparing the results of the exponential tagger to the maximum entropy tagger when there is only a limited amount of data available. The results are summarized in the Table 10. Since the public version of the MaxEnt tagger cannot be modified to take the advantage of the neither the *mixed* nor *independent* dictionary, we have compared it only to the *automatic* dictionary version of the exponential tagger. To save space, the results are tabulated only for the training data sizes of 2000, 5000 and 20000 words. Again, only the "true" word error rate is reported.

As the tables show, for the languages we tested, the exponential, feature-based tagger we adapted from (Hajič and Hladká, 1998) achieves similar results as the Maximum Entropy tagger[14] [15]. (using exactly the same (**full**) training data; the "score" is 3:3, with the MaxEnt tagger being substantially better on English; probably the development lan-

[14]Otherwise the acknowledged leader in English tagging

[15]We noticed only substantial difference in tagging speed. The runtime speed of the MaxEnt tagger is lower, only about 10 words per second vs. almost 500 words per second; it should be noted however that we are comparing MaxEnt's java bytecode and C.

Table 2: Results (Error rate, ER) on full training data, only true words counted (no punctuation)

Dictionary:	Automatic		Mixed		Independent	
Language	Baseline	Full	Baseline	Full	Baseline	Full
English	11.42%	9.18%	11.40%	7.91%	7.07%	3.58%
Czech	23.02%	18.83%	22.61%	14.78%	19.40%	9.59%
Estonian	16.12%	13.95%	16.19%	12.98%	9.94%	5.34%
Hungarian	8.35%	8.16%	8.31%	8.00%	3.55%	2.58%
Romanian	10.87%	7.76%	10.81%	7.34%	7.49%	3.35%
Slovene	20.53%	16.26%	20.01%	13.29%	17.29%	9.00%

Table 3: Error rate on reduced training data, dictionary: automatic

Language	Training data size					
	1000	2000	5000	10000	20000	Full
English	36.20%	29.36%	23.47%	18.27%	14.46%	9.18%
Czech	48.22%	42.95%	36.54%	30.97%	27.08%	18.83%
Estonian	48.14%	42.10%	32.44%	26.81%	21.51%	13.95%
Hungarian	39.68%	32.21%	23.94%	18.04%	13.92%	8.16%
Romanian	40.61%	35.02%	25.06%	19.26%	15.16%	7.76%
Slovene	45.84%	39.58%	33.12%	28.60%	24.50%	16.26%

Table 4: Error rate on reduced training data, dictionary: mixed

Language	Training data size					
	1000	2000	5000	10000	20000	Full
English	36.15%	29.58%	22.93%	17.70%	14.00%	7.91%
Czech	48.97%	41.93%	34.37%	28.10%	23.31%	14.78%
Estonian	48.24%	42.79%	32.98%	26.60%	21.02%	12.98%
Hungarian	39.87%	32.71%	23.63%	17.98%	13.82%	8.00%
Romanian	42.85%	35.70%	25.46%	19.23%	14.81%	7.34%
Slovene	46.74%	39.88%	32.00%	26.20%	21.73%	13.29%

Table 5: Error rate on reduced training data, dictionary: "independent"

Language	Training data size					
	1000	2000	5000	10000	20000	Full
English	10.29%	7.64%	5.53%	4.54%	3.83%	3.58%
Czech	22.51%	18.07%	17.33%	15.10%	12.62%	9.59%
Estonian	13.11%	11.95%	10.70%	9.29%	8.10%	5.34%
Hungarian	6.84%	5.35%	4.29%	4.07%	3.48%	2.58%
Romanian	13.11%	9.47%	7.81%	6.18%	5.07%	3.35%
Slovene	24.63%	19.17%	16.17%	14.12%	12.62%	9.00%

guage bias shows here[16]). However, when the training data size goes down, the advantage of predicting the single morphological categories separately favors the exponential tagger (with the notable and substantial exception of English). The less data, the larger the difference (Tab 10).

[16]On the other hand, the Exponential tagger has been developed on Czech originally and it lost on this language. It should be noted that the original version of the exponential tagger did contain Czech-specific features, removed here, which might in fact do better.

The resulting accuracy (of both taggers) is still unsatisfactory not only from the point of view of results obtained on English, but also from the practical point of view: approx. 85% accuracy (Czech, Slovene) typically means that about five of each six sentences of length 10 words has at least one error in it. That is bad news e.g. for parsing projects involving tagging as a preliminary step.

Table 6: POS Error rate on reduced training data, dictionary: automatic

Language	Training data size					
	1000	2000	5000	10000	20000	Full
English	26.77%	20.82%	16.11%	11.86%	9.48%	5.64%
Czech	24.32%	20.20%	13.46%	9.70%	7.22%	3.72%
Estonian	35.81%	30.52%	23.02%	18.26%	14.31%	8.46%
Hungarian	30.54%	24.99%	18.09%	13.15%	10.29%	5.81%
Romanian	31.33%	27.59%	19.24%	14.51%	11.25%	5.21%
Slovene	27.16%	23.15%	17.01%	12.89%	9.74%	5.61%

Table 7: POS Error rate on reduced training data, dictionary: mixed

Language	Training data size					
	1000	2000	5000	10000	20000	Full
English	26.69%	21.09%	15.82%	11.53%	9.08%	4.94%
Czech	24.32%	20.61%	13.47%	10.19%	7.37%	3.76%
Estonian	36.48%	31.76%	23.55%	18.21%	14.32%	8.20%
Hungarian	30.28%	25.25%	17.59%	12.89%	10.15%	5.64%
Romanian	33.56%	28.34%	20.03%	14.52%	11.03%	5.04%
Slovene	27.58%	23.30%	16.85%	12.59%	9.88%	5.12%

Table 8: POS Error rate on reduced training data, dictionary: "independent"

Language	Training data size					
	1000	2000	5000	10000	20000	Full
English	6.42%	5.36%	3.63%	3.02%	2.53%	2.43%
Czech	3.21%	2.85%	2.17%	2.01%	1.65%	1.12%
Estonian	6.71%	6.32%	5.27%	4.31%	3.77%	2.36%
Hungarian	5.35%	4.42%	3.39%	3.18%	2.75%	2.04%
Romanian	9.51%	6.54%	5.36%	4.00%	3.18%	1.89%
Slovene	6.10%	5.19%	4.04%	3.59%	3.25%	2.08%

5 Conclusions

5.1 The Differences Among Languages

The following discussion abstracts from the tagset design, relying on the fact that the Multext-East project has been driven by a common tagset guidelines to an unprecedented extent, given the very different languages involved. At the same time, we acknowledge that even so, their design for the individual languages might have influenced the results. Also, the quality of the annotation is an important factor; we believe though that the late data we obtained for the experiments described here are within the range of usual human error and do not suffer from negligence[17].

First of all, it is clear that these languages differ substantially just by looking at the simple training data statistics, where the number of unique tags seen in a relatively small collection of about 100k tokens is high - from 401 (Hungarian) to 1033 (Slovene); compare that to English with only 139 tags. It is however interesting to see that the average per-token ambiguity is much more narrowly distributed, and in fact English ranks 3rd (after Hungarian and Slovene), Czech being the last with almost every other token ambiguous on average. This ambiguity does not correspond with the results obtained: Slovene, being the second least ambiguous, is the second most difficult to tag. Only Czech behaves consistently by tailing the pack in both cases.

5.2 Comparison to Previous Results

Any comparison is necessarily difficult due to different evaluation methodologies, even within the "best-only", accuracy-based reporting. Nevertheless, we will try.

For Romanian, Tufis in his recent work (Tufis, 1999) reports 98.5% accuracy (i.e. 1.5% error rate) on Romanian, using the classifier combination approach advocated by e.g. (Brill and Wu, 1998). His

[17] Specifically, we are sure that the post-release Czech, Slovene and Hungarian data we are using are without annotation defects beyond the usual occasional annotation error, as they have been double checked, and we also believe that the other two languages are reasonably clean. Bulgarian, although present on the CD, is unfortunately unusable since it has not been manually annotated; for English, see above.

Table 10: Error rate comparison on reduced training data, automatic dictionary

	Training data size					
	2000		5000		20000	
Language	ME	Exp	ME	Exp	ME	Exp
English	26.03%	29.36%	17.70%	23.47%	9.61%	14.46%
Czech	50.77%	42.95%	41.95%	36.54%	28.16%	27.08%
Estonian	51.08%	42.10%	40.09%	32.44%	25.50%	21.51%
Hungarian	41.12%	32.21%	30.68%	23.94%	17.27%	13.92%
Romanian	42.88%	35.02%	30.07%	25.06%	16.67%	15.16%
Slovene	49.46%	39.58%	39.34%	33.12%	27.77%	24.50%

results are well above the 3.29% error rate achieved here (with even a larger tagset of 1391 vs. 486 here), but the paper does not say how this number has been computed (training data size, the all-token/words-only question) thus making any conclusions difficult to make. He also argues that his method is language independent but no results are mentioned for other languages.

For Czech, previous work achieved similar results (6.20% on newspaper text using the all-tokens-based error rate computation, on 160,000 training tokens; vs. 7.04% here on approx. half that amount of training data; same handling of unknown words). This is in line with the expectations, since the same methodology (tagging as well as evaluation) has been used, except the features used in that work were specifically tuned to Czech.

The most detailed account of Slovene (Erjavec et al., 1999) reports various results, which might not be directly comparable because it is unclear whether they use the all-tokens-based or words-only computation of the error rate. They report 6.421% error rate on the full tagset on *known* words, and 13.583% on all words (tokens?) incl. unknown words (the exp. tagger we used achieved 13.82% on all tokens, 16.26% on words only). They use almost the same data (Orwell's 1984, but leaving out the Appendices)[18]. They also report that the original Czech-specific exponential tagger used as a basis for the work reported here achieved 7.28% error rate on Slovene on full tags on the same data, which means that by the changes to the exponential tagger aimed at its language independence we introduced in Section 3, we have not achieved any improvement (on Slovene) of the exp. tagger (the error rate stayed at 7.26% – using all-tokens-based evaluation numbers, dictionary available; but the data was not exactly the same, presumably).

5.3 Dictionary vs. Training Data

This is, according to our opinion, the most interesting result of the experiments described so far. As already Table 2 clearly suggests, even the baseline tagging results obtained with the help of an independent dictionary are comparable (if not better) than the fully-trained tagger on 100k words, but without the dictionary information. The situation is even clearer when comparing the POS-only results: here the "independent" dictionary results are better by far, with almost no training data needed.

Looking at the characteristics of the languages, it is apparent that the inflections cause the problem: the coverage of a previously unseen text is inferior to the usual coverage of English or another analytical language. Therefore, unless we can come up with a really clever way of learning rules for dealing with previously unseen words, it is clearly strongly preferable to work on a morphological dictionary[19], rather than to try to annotate more data.

6 Future Work

We would like to compare more taggers using still other methodologies, especially the MBT tagger, which achieved the best results on Slovene but which was not available to us at the time of writing of this paper, and obviously, try to use the classifier combination method on them, to confirm the really surprisingly good results on Romanian and test it on the other languages as well.

We would also like to enrich the best taggers available today (such as the Maximum Entropy tagger) to use the dictionary information available and compare the results with the Exponential Feature-based tagger we have been using in the experiments here.

For Czech and Slovene, the results are still far below what one would like to see (in absolute terms). It seems that the key lies in the initial feature set definition - including statistical tagset clustering, which might potentially lead to more reliable estimates of certain parameters while using still the same size of training data.

[18]Their tag count is lower (1021) than here (1033), but that's not really relevant. They do not report the average ambiguity or a similar measure.

[19]Not necessarily manually - apparently, even a partially supervised method would be of tremendous help.

7 Acknowledgements

The author wishes to thank to many MULTEXT-East participants for their efforts to improve the original data, especially to Niki Petkevič, Tomaz Erjavec, Heiki-Jaan Kaalep and Gábor Prószéky, and for providing the final versions of the annotated data for the experiments. Any errors and mistakes are solely to be blamed on the author, not the annotators, of course.

References

Eric Brill and Jun Wu. 1998. Classifier combination for improved lexical disambiguation. In *Proceedings of ACL/COLING'98*, pages 191–195, Montreal, Canada. ACL/ICCL.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–565.

Jean-Pierre Chanod. 1997. Current developments for Central & Eastern European languages. In *Proceedings of EU Project meeting TELRI I*, Romania.

Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas. ACL.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger generator. In *Proceedings of WVLC 4*, pages 14–27. ACL.

Tomaz Erjavec, Saso Dzeroski, and Jakub Zavrel. 1999. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. Technical Report IJS-DP 8018, Dept. for Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia, April 2nd.

Jan Hajič and Barbora Hladká. 1997. Tagging of inflective languages: a comparison. In *Proceedings of ANLP'97*, pages 136–143, Washington, DC. ACL.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of ACL/COLING'98*, pages 483–490, Montreal, Canada. ACL/ICCL.

Andrew W. Mackie Julian Benello and James A. Anderson. 1989. Syntactic category disambiguation with neural networks. *Computer Speech and Language*, 3:203–217.

Bernard Merialdo. 1992. Tagging text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of EMNLP 1*, pages 133–142. ACL.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, England.

Dan Tufis. 1999. Tiered tagging and combined language models classifiers. In *Proceedings of Text, Speech and Dialogue'99*, Mariánské Lázně, Czech Republic, Sept. 15–18.

Jean Véronis. 1996a. Multext-East (Copernicus 106). http://www.lpl.univ-aix.fr/projects/multext-east.

Jean Véronis. 1996b. Multext-East language-specific resources (Copernicus 106). http://www.lpl.univ-aix.fr/projects/multext-east/MTE2.html.