Bottom–Up Tagset Design from Maximally Reduced Tagset*

Péter Dienes and **Csaba Oravecz**
Research Institute for Linguistics
Hungarian Academy of Sciences, Budapest
{dienes,oravecz}@nytud.hu

Abstract

For highly inflectional languages, where the number of morpho-syntactic descriptions (MSD) is very high, the use of a reduced tagset is crucial for reasons of implementation problems as well as the problem of sparse data. The standard procedure is to start from the large set of MSDs incorporating all morphosyntactic features and design a reduced tagset by eliminating the attributes which play no role in disambiguation. This paper presents the opposite approach which using a greedy algorithm maximally reduces a tagset without loss of information, and instead of elimination, re-introduces features. This process can arrive at a very small tagset and result in accuracy comparable to that achieved with larger tagsets designed by elimination. The language model based on the reduced tagset needs fewer parameters and training time decreases significantly.

1 Introduction

In highly inflectional languages, the number of morpho-syntactic descriptions (MSD), required to descriptionally cover the content of a word-form lexicon, tends to rise quite rapidly, approaching a thousand or even more set of distinct codes. For the purpose of automatic disambiguation of arbitrary written texts, using such large tagsets would raise very many problems, starting from implementation issues of a tagger to work with such a large tagset to the more theory-based difficulty of sparseness of training data. Tiered tagging (Tufiş, 1998) is one way to alleviate this problem by reformulating it in the following way: starting from a large set of MSDs, design a reduced tagset, Ctag-set, manageable for the current tagging technology. The standard procedure is to start from the large set of MSDs incorporating all morphosyntactic features and design a reduced tagset by eliminating the attributes which play no role in disambiguation. However, there are a number of reasons for which one can question whether such a process can produce anything close to an optimal tagset and eliminate all irrelevant features. In section 2 we briefly outline these reasons and in section 3 present the data used. Section 4 suggest and alternative approach that just takes the opposite way and a maximally reduced tagset as starting point for the design process. Section 4.2 will present some preliminary results on tagging accuracy and error analysis comparing the performance of the tagging process with tagsets of different cardinality. Conclusions and suggestions for further work will follow in section 5.

2 Tagset design and highly inflected languages

The combinatorial possibilities of inflection and derivation in highly inflectional languages pose a challenge for corpus annotation in that it is difficult to establish a set of morphosyntactic descriptions that does justice to the rich morphosyntactic information encoded within the words and at the same time remains computationally tractable. The design process of a reduced tagset has to consider two fundamental requirements: to identify and leave out the features/values in the MSDs which do not provide relevant clues for the contextual disambiguation, and to make it possible to recover as accurately and fast as possible the information eliminated in the previous phase.

The standard approach is usually a trial-and-error one augmented by some algorithm and re-

* The research reported in this paper was supported by grant number T026091 of the Országos Tudományos Kutatási Alap.

lies both on human introspection and evidence provided by the data analysis (Elworthy, 1995), (Chanod and Tapanainen, 1995), (Tufiş, 2000). One can use an information loss-less algorithm to convert the MSD-set into a Ctag-set which might reduce the size of the tagset with 10-20% (Brants, 1995); however, this is too little for a large initial tagset. Modifying such an algorithm to allow for limited ambiguity (that is losing a limited amount of information), could result in a drastic reduction of the Ctag-set, up to a cardinality which is within the restrictions imposed by the available training data and computing power (Tufiş, 1998). Nevertheless, this procedure fails to obtain the optimum result for at least two reasons: there is, even if limited, loss of information and the recoverability of information contained in the original MSDs is not preserved; and features that do not appear in ambiguity classes are usually not submitted to the reduction algorithms and may be preserved unnecessarily.

3 Data analysis

The language resource of our analysis consisted of the whole current stock of the Hungarian National Corpus (approximating 80m words) compiled into a word frequency list as input to the morphological analysis. The initial assumption is that this large number of word forms contain all possible ambiguity classes that can occur in the language. Table 1 presents some basic statistics on the range of word form variation found in the corpus.

Entries	Word forms	Lemmas
74,063,211	1,728,771	429,612[1]

Table 1: The distribution of word forms

The word form list was processed with the morphological analyzer developed originally for Hungarian (Prószéky and Tihanyi, 1996). An MSD notation was constructed which represented the POS category and the inflectional structure of the word, and which can in principle be mapped into the EAGLES compliant encoding scheme developed in Multext–East (Erjavec and Monachini, 1997). The MSD scheme, as an initial step in tagset creation, was converted into an attribute/value single string representation. The intent at this stage was merely to preserve in a concise and consistent notation all the information provided by the MSD that is relevant for tagging. Table 2 displays the features encoded in this initial Ctag scheme (Full set) for the major POS categories.

As the cardinality of the full initial tagset was too high to be handled by current tagging methods (2148), especially by statistical taggers, a medium tagset was designed by feature elimination as detailed in (Tufiş et al., 2000). (This medium tagset does not ensure full recoverabilty, though.) These two tagsets serve as the basis of comparison in the evaluation of the alternative approach for tagset creation we will propose below. In the experiments, two HMM taggers are used: Thorsten Brants' trigram TnT tagger (Brants, 1998) and the MULTEXT-ISSCO (M–I) bigram tagger (Gilbert and Amstrong, 1995) used in the Multext-East project (Erjavec and Ide, 1998). The training corpus consists of two register-diverse corpora: the first three quarters of Orwell's 1984 and newspaper text, adding up to 87969 tokens altogether. The test corpus includes the rest of the Orwell and newspaper texts, 21267 tokens in total. The MULTEXT-ISSCO tagger is trained with the Baum-Welch algorithm. The TnT tagger has the problem of learning possible ambiguity classes and words from the training corpus only. To remedy this situation, after the training phase, we enriched the generated lexicon file with further ambiguities and added words from the test corpus with their ambiguity classes. The tagging results with the above tagsets are presented in Tables 3 and 4.

	Error perc.	Error rate	Perf.
M-I	23.83%	6.04%	93.96%
TnT	14.00 %	3.55%	96.45%

Table 3: Full tagset (2148 tags)

[1] The number of lemmas were calculated on the assumption that alternatives in ambiguous cases were evenly distributed. This is obviously false but the correct figure could only be arrived at after the corpus has been completely disambiguated.

POS	Num	Pers	Stem [NAR] Mood/Tense [V]	Case [N] Def [V]	Owner's Num	Owner's Pers	Total
N	2 [PS]	3 [123]	5 [QAVNP]	21	2 [PS]	3 [123]	2058*
A			2 [AV]				2*
R			2 [RV]				2*
V	2 [PS]	3 [123]	5 [PRCSI]	3 [ID2]			79*
Invariant minor categories: Q, D, Z, RP, C, Int, Y							7
							2148

N = Noun A = Adjective R = Adverb V = Verb
Q = Numeral D = Article Z = Verbal prefix RP = Postposition
C = conjunction Y = Abbreviation Int = Interjection
Def = Agreement in definiteness with object (def, indef, 2nd person)
Owner's Num = sing. or plural owner Owner's Pers = person marker of owner
* = not all combinations are possible, so not a simple product
[NAR][V][N] = POS categories to which the attributes apply

Table 2: The initial Ctag scheme (F set)

	Error perc.	Error rate	Perf.
M-I	22.48%	5.70%	94.3%
TnT	13.37%	3.39%	96.61%

Table 4: Medium tagset (240 tags)

4 Maximal reduction and bottom–up design

4.1 Maximal reduction of the tagset

The alternative approach using a greedy algorithm maximally reduces a tagset without loss of information, and instead of elimination, reintroduces features. This process first arrives at a very small tagset and the application of this tagset in tagging results in a dramatic drop in accuracy compared to that achieved with a tagset designed from MSD reduction with the elimination algorithm in (Tufiş, 2000) and linguistic introspection. However, even the reintroduction of a few morphosyntactic features leads to a sharp increase in accuracy comparable to that achieved with larger tagsets designed by elimination. The language model based on the reduced tagset needs fewer parameters and training time decreases significantly.

The construction of the minimal tagset proceeds the following way. First a graph G is established whose vertices are the tags of the initial tagset. Two points (tags) are connected with an edge if and only if there exists a word which can be assigned both tags. That is, two tags are not connected if they do not occur in an ambiguity class. Then, a partition of this graph is created as follows:

> x and y are in the same partition if and only if there is no (x, y) edge.

The problem is equivalent to the colourability problem of the graph G:

Colourability problem: The aim is to colour the vertices of a graph G with as few colours as possible so that neighbouring vertices have different colours.

In the general case the problem of finding the minimal number of colours (*chromatic number*, $\chi(G)$) cannot be solved within polynomial time. Nevertheless, certain estimations of $\chi(G)$ can be given. The algorithm to be discussed and applied here, for example, yields the result:

$$\chi(G) \leq 1 + \max_{g \in G} \phi(g),$$

where $\phi(g)$ is the degree of the point g, i.e. the number of its neighbours. The algorithm is a simple greedy algorithm. The colours are nonnegative integers.

Algorithm:

1. *Ordering phase*: order the vertices of the graph in any way;
2. *Colouring phase*: for each $i = 1, 2, \ldots$ colour the ith vertex with the smallest available colour. Make this colour unavailable for all neighbouring vertices.

In fact, according to *Brooks-theorem* the *chromatic number* can easily be decreased by one, i.e. (Gross and Yellen, 1998):

Brooks-theorem.

$$\chi(G) \leq \max_{g \in G} \phi(g)$$

Now, consider the graph obtained from the 74-million-word wordlist, tagged with the full tagset. Out of the 1105 tags 968 occur in ambiguity classes, the maximal degree of the vertices of the graph is 192. According to the above theorems, this means that the tagset can be reduced to 192 tags without merging ambiguity classes. This in itself is quite a considerable decrease in the number of tags.

However, for graphs containing several vertices, the estimations obtained from these theorems might lie far over the actual value of the *chromatic number*. This might especially be the case if we deal with graphs obtained from natural language corpora, because these graphs seem to be unsaturated. Figure 1 presents the top 20 degrees of the "Hungarian graph".

The data clearly shows that there are two vertices with fairly large number of points, but the degree of vertices decreases rapidly. This might suggest that this graph can be coloured with relatively few colours. Indeed, the actual experiment with the algorithm described above yielded a surprising result: the graph can be coloured with 10 colours, that is, the number of tags can be reduced to 10 without merging ambiguity classes and retaining full recoverability.

4.2 Enriching the minimal tagset

The minimal tagset containing only 10 tags significantly reduces the problem of sparse data. However, with the radical reduction of the tagset, though recoverability is retained, we have lost important environmental information which could serve as tagging clues for the tagger. Thus, as illustrated in Tables 3 and 5, we face radical decrease in tagging accuracy even with respect to the results exhibited by the full tagset. (The cardinality of the tagset is indicated in parentheses.)

1	NS3NN	192
2	R	184
3	VS3RI	71
4	P	57
5	NS3NA	54
6	AS_A	54
7	RP	49
8	NP3NN	47
9	NS3NP	39
10	NS3NS	36
11	NS3PC	36
12	NS3NNS3	36
13	AS_V	35
14	NS3ND	32
15	NS3NI	31
16	NS3N2	31
17	Z	29
18	NS3NX	29
19	NS3NT	28
20	NS3N3	28

Figure 1: Degree of vertices

	Error perc.	Error rate	Perf.
M-I	32.78%	8.31%	91.69%
TnT	60.94%	15.45%	85.55 %

Table 5: Minimal tagset (10)

The inaccuracy originating from the minimal tagset is especially spectacular in the case of the HMM-based trigram TnT tagger. Here, 15.54% of all words is mistagged, which is over 60% error on ambiguous words. The decrease in the actual performance of the MULTEXT-ISSCO tagger is less conspicuous, though still significant.

One important problem with the minimal tagset is that it fails to indicate punctuation, that is, punctuation tags (CPUNCT, OPUNCT, SPUNCT and WPUNCT) are merged with each other and several other tags. The increase in the performance of the TnT tagger is significant if these four tags are retained. This is illustrated in Table 6.

Interestingly, the reactions of the MULTEXT-ISSCO tagger to this small change is less radical:

	Error perc.	Error rate	Perf.
M-I	31.81%	8.06%	91.94%
TnT	18.81%	4.77%	95.23%

Table 6: Minimal tagset with punctuation tags (14)

the bigram HMM-base tagger seems to depend less on the information provided by punctuation tags. One possible reason for the difference of the behavior between the two models can be that information before the punctuation mark is unavailable for the bigram tagger, regardless whether it "knows" that the word to be disambiguated is preceded by a punctuation mark. On the other hand, a trigram tagger can "learn" to disregard punctuation tags and consider the previous tags only. Whether this assumption can emprically be justified, however, is subject to careful future research.

Another clue that can help the proper identification of tags is the distribution of the main categories, i.e. nouns, verbs and adjectives. This type of information is especially useful for the bigram tagger, for reasons discussed above (cf. Table 7).

	Error perc.	Error rate	Perf.
M-I	19.66%	4.98%	95.02%
TnT	14.84%	3.76%	96.24%

Table 7: Minimal tagset NAV heads only (30)

As we can see, this information provided by the re-introduction of the main head categories proves to be crucial for the trigram tagger as well. Note that the performance of the MULTEXT-ISSCO tagger with these 30 tags is higher than the performance with the handcrafted, "linguistically motivated" medium tagset.

The combination of the two types of information does not increase the performance of the tagger significantly. Similarly, with the re-introduction of all head categories, the error of the taggers does not decrease crucially, as is illustrated in Table 8.

Hungarian has a very rich case system with 22 cases, which might offer important tagging clues in the disambiguation process. In the experiment, in order to avoid the proliferation of tags, we reduced the possible morphological cases to three: nominative, accusative and other case. The results thus obtained are of considerable importance: though the performance of the bigram taggers decreases insignificantly, the trigram tagger's performance reaches the performance shown with the hand-crafted medium tagset.

	Error perc.	Error rate	Perf.
M-I	18.94%	4.80%	95.2%
TnT	14.35%	3.64%	96.36%

Table 8: Minimal tagset with all head categories (39)

	Error perc.	Error rate	Perf.
M-I	19.12%	4.85%	95.15%
TnT	13.54%	3.43%	96.57%

Table 9: Minimal tagset with head cat.s and N case (59)

However, these results are only preliminary inasmuch as only considerably larger training and test corpora and much more extensive testing could provide reliable justification for the re-introduction of one or the other features. Still, these preliminary experiments indicate that a bottom-up procedure can perform at a similar level to a top-down eliminative approach.

5 Conclusion

The paper described a method of maximally reducing a tagset which is supplemented by a "bottom–up" procedure of re-introduction of features, which can achieve acceptable tagging accuracy using a very small tagset with full MSD recoverability. This method is based on a fast and effective algorithm and not only leads to building a language model with fewer parameters in a comparably shorter training time but could also give insight to finding those morphosyntactic features that provide relevant information as contextual clues in ambiguity resolution.

Further investigation should involve more types of taggers including a rule based application (Alexin et al., 1999) as well. It would also be interesting to see how far tagging performance can be improved by this method[2], and extend the experiments to other languages where the MSD cardinality and the size of the tagset used in tagging experiments is high (Harris et al., 2000), (Hajič and Hladka, 1998). Another crucial advantage lies in the possibility of algorithmic feature re-introduction, the problem of which should also be addressed in the future.

References

Zoltán Alexin, Tamás Váradi, Csaba Oravecz, Gábor Prószéky, János Csirik, and Tibor Gyimóthy. 1999. Fgt – a framework for generating rule-based taggers. In *ILP-99 Late-Breaking papers*, Bled, Slovenia.

Thorsten Brants. 1995. Tagset reduction without information loss. In *Proceedings of ACL–95*, Cambridge, MA.

Thorsten Brants, 1998. *TnT – A Statistical Part-of-Speech Tagger, Instalation and User Guide.* University of Saarland.

Jean-Pierre Chanod and Pasi Tapanainen. 1995. Creating a tagset, lexicon and guesser for a French tagger. In E. Tzoukermann and S. Armstrong, editors, *From Texts to Tags: Issues in Multilingual Language Analysis: Proceedings of the ACL SIGDAT Workshop*, pages 58–64, Geneva.

David Elworthy. 1995. Tagset Design and Inflected Languages. In *Proceedings of the ACL-SIGDAT Workshop*, Dublin. (also available as cmp-lg/9504002).

Tomaž Erjavec and Nancy Ide. 1998. The MULTEXT-EAST corpus. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation, LREC'98*, pages 971–974, Granada. ELRA.

Tomaz Erjavec and M. Monachini. 1997. Specifications and notation for lexicon encoding. COP Project 106 Multext-East, Deliverable D1.1 F (Final Report).

R. Gilbert and S. Amstrong. 1995. Tagging tool. MULTEXT Deliverable 2.4.1.

Jonathan Gross and Jay Yellen. 1998. *Graph Theory and Its Applications.* CRC Press.

Jan Hajič and Barbora Hladka. 1998. Tagging inflective languages: prediction of morphological categories for a rich structured tagset. In *Proceedings of the 36th annual meeting of the ACL – COLING*, Montreal, Canada.

Papageorgiou Harris, Prokopidis Prokopis, Giouli Voula, and Piperidis Stelios. 2000. A unified pos tagging architecture and its application to Greek. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens.

Gábor Prószéky and László Tihanyi. 1996. Humor – a Morphological System for Corpus Analysis. In *Proceedings of the first TELRI Seminar in Tihany*, pages 149–158, Budapest.

Dan Tufiş, Péter Dienes, Csaba Oravecz, and Tamás Váradi. 2000. Principled hidden tagset design for tiered tagging of Hungarian. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens.

Dan Tufiş. 1998. Tiered tagging. Technical Report 32, RACAI.

Dan Tufiş. 2000. Using a large set of eagles-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens.

[2]Present tagger implementations cannot produce above around 96% for Hungarian, which constitutes an actual limit for testing this method.