Korpusar i forskning och undervisning • Corpora in Research and Teaching. Rapport från ASLA:s höstsymposium, Växjö 11–12 november 1999 • Papers from the ASLA symposium Corpora in Research and Teaching, Växjö, 11–12 November 1999. Utg. av Gunilla Byrman, Hans Lindquist, Magnus Levin. Uppsala: ASLA. 2000. 19–31.

Enhancing tagging performance by combining knowledge sources[1]

Lars Borin
Uppsala University

1 Introduction

The topic of this paper is an ongoing effort to exploit combinations of existing natural language processing (NLP) resources in order to reach part-of-speech (POS) tagging performance in excess of that which any single resource is able to provide.

The context of the effort is the ETAP project, a parallel translation corpus project funded by the Bank of Sweden Tercentenary Foundation. The aim of the project is to create an annotated and aligned multilingual translation corpus which will be used as the basis for the development of methods and tools for the automatic extraction of translation equivalents for applications such as machine translation systems.

To this end, we are investigating to which extent it is possible to reuse existing – meaning either developed in our department in some other context, or freely available on the WWW – NLP resources for the task of tagging the languages of the project. As a general rule, we may say that the amount of such resources is growing quite fast at the present time. On the other hand, their availability is highly dependent on the language, from almost unlimited numbers for English,

[1] The research reported in this paper was carried out within the ETAP (Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter; in English: "Creating and annotating a parallel corpus for the recognition of translation equivalents") project, supported by the Bank of Sweden Tercentenary Foundation as part of the research programme Translation and Interpreting – a Meeting between Languages and Cultures. See http://www.translation.su.se/

over a few different POS taggers for German or Swedish[2], to practically nothing for a language like Polish[3].

Even in the cases where more than one tagger is available, their performance on our corpus may be fairly uneven, since they represent different tagger technologies, come with lexicons and tagsets of different size, and have been trained on different types and amounts of text. However, this can be used to advantage, since it seems that systematic differences between taggers can be exploited to enhance tagging performance.

Another thread of investigation in the project deals with the relationship between POS tagging and word alignment. Since we are working with parallel translation corpora, we are investigating the possiblity of using word alignment to complement tagging. This is achieved by taking advantage of systematic part-of-speech correspondences between languages, so that a higher-precision tagger for language A – e.g. Swedish – may correct and complement the lower-precision (or nonexistent) tagging of a parallel text in language B – e.g. Polish – with which it has been aligned at the word level.

Both these efforts represent a recycling of the *knowledge* embodied in existing resources, rather than merely the straightforward reuse of those resources, in a narrower sense of the word, and we now turn to the – admittedly not completely sharp – distinction between the two kinds of reuse.

2 Reusing knowledge in computational linguistics

In language engineering, just as in software development in general, reusability is often equated with *modularity*. Modularity in turn presupposes standardisation, since the modules cannot communicate other than through a mutually agreed-upon – i.e., standardised – inter-

[2] In addition, the tagged corpora which are used to train POS taggers are still very few in number, so that e.g. Swedish taggers, regardless of their provenance or the tagging technology used, tend to be trained on the SUC corpus (Ejerhed and Källgren 1997).

[3] In recent NLP terminology, this is the difference between high-density and low-density languages.

face. However, the internal workings of the modules still need not be subject to standardisation in this way[4].

The development of a general linguistic resource for NLP is a major undertaking, and it is thus natural that there are various ongoing standardisation efforts in the language engineering community, e.g. the Text Encoding Initiative (TEI) for the markup of linguistic resources, EAGLES for content models for different kinds of such resources (Godfrey and Zampolli 1997), and GATE (Cunningham *et al.* 1995) for a standardised environment in which NLP modules can be combined in various ways.

2.1 Combining knowledge

Standardisation, or, rather, commensurability, is a prerequisite for the more specific kind of reuse discussed here, namely the reuse of the *knowledge* embodied in existing linguistic resources, in ways which were not foreseen when the resources were created.

Another prerequisite is that the knowledge sources be (at least in part) *complementary*, i.e., there is no point in combining, e.g., part-of-speech taggers which make the same errors, or where the errors of one tagger is a proper subset of those of the other one.

In our view, it is worthwhile to attempt such a combination of knowledge sources, since each of them is *incomplete*, i.e. there are no perfect taggers, all-encompassing lexicons, etc., at least not for general language.

Here, we will look at two kinds of knowledge combination relevant for the larger endeavour of annotating a multilingual parallel corpus for enabling the extraction of translation equivalents from it:

---

[4] Although it would seem that a standardised interface will impose some limitations on the kinds of representations that can be internally manipulated, in practice this is not a great problem. In the physical world, the dimensions of a conduit will inevitably limit the size of objects which are meaningful to handle in activities linked up by this conduit. This is because, in the physical world, for all practical purposes, the whole often cannot be restored from the parts; you cannot cut up a person, send the pieces through e.g. a tube mail system, and expect to be able to put the person together again at the other end. With information, however, this is fully possible, so internal representations can be arbitrarily larger than the pieces that can pass through the interface (although these pieces themselves – putting it in a somewhat simplified way – cannot both be arbitrarily small and arbitrarily ordered).

(1) The combination of several off-the-shelf part-of-speech taggers;

(2) The combination of a part-of-speech tagger with word alignment

The first endeavour has precedents, both in computational linguistics and outside it. In the machine learning community, the idea of combining classifiers – e.g., neural networks trained on the same classification task – for enhancing accuracy, is an old one, going back at least to the mid-sixties (Tumer and Ghosh 1999). Several regimes for classifier combination have been proposed, from simple averaging, over majority voting and more complicated non-linear models, to training a new classifier on the basis of the combination. All these methods have in common that they are *knowledge-poor*, i.e. they require no domain knowledge for their implementation. With other such methods, they share the need for relatively large amounts of training data, and the feature of being supervised methods, i.e. the 'right' answer must be part of the training data.

POS taggers are classifiers in this sense, and it is natural to see how the methods developed for general machine learning could be applied for this specific machine learning task as well. The experiments with POS tagger combination which have been reported in the literature (Màrquez *et al.* 1998; Brill and Wu 1998; van Halteren *et al.* 1998) have all adhered faithfully to this kind of knowledge-poor, supervised training regime. To my knowledge, the work reported here represents the first attempt to apply a *knowledge-rich* method to the problem of combining POS taggers, by formulating linguistically motivated rules for how tagger differences should be utilised in the combination of taggers.

2.2 Combining part of speech taggers

2.2.1 Step 1: Finding taggers

The first step in the tagger comparison procedure was the procurement of taggers to compare. Here, I will discuss the comparison of German taggers, but the procedure described is quite independent of

language[5]. For German, we found three publicly available part-of-speech (POS) taggers, *Morphy* (Lezius *et al.* 1998), *QTAG* (Mason 1997), and *TreeTagger* (Schiller *et al.* 1995).

2.2.2 Step 2: Evaluating the taggers

The evaluation of the taggers was carried out according to the following procedure. One or two short texts from the various subcorpora of the ETAP project were tagged with each of the taggers. Ten sentences were then picked out and the number of correct and incorrect tags in them counted.

Of the three German taggers evaluated, one, QTAG, turned out to have unacceptably low accuracy[6]. This was probably due to it having been trained on nineteenth century fiction (Oliver Mason, *p.c.*), while the ETAP texts are contemporary non-fiction.

The tagsets of the two remaining taggers differ considerably in size. TreeTagger tags encode mainly part of speech, but no inflectional information, (or at the most very coarse-grained inflectional distinctions, e.g. finite vs. infinite verb forms), while Morphy tags represent richer morphosyntactic desriptions.

In Table 1, the performance of the two German taggers is shown for two text types, technical manuals from the Scania subcorpus, and political prose from the German translation of the Swedish *Statement of Government Policy* (SGP) of 1988 and 1996. Accuracy percentages are calculated as: CORRECTLY TAGGED TOKENS/ALL TOKENS.

Table 1: Tagger accuracies

Tagger/tagset	*Scania*	SGP
TreeTagger	96.3%	96.2%
Morphy/full	90.4%	93.8%
Morphy/reduced	94.7%	95.4%

[5] Apart from such obvious considerations as the availability of computational resources for a particular language, of course. Thus, for English, our search for freely available resources turned up three taggers with altogether 10 different tag sets to choose among, while we have not been able so far to find even a single tagger for Polish.

[6] We set the accuracy threshold for inclusion in the comparison experiment at 90%, since this seems to be the commonly acknowledged chance baseline for POS tagging – i.e. the accuracy that would result if the most probable tag would be assigned to each word, regardless of context – at least for English (see, e.g., Voutilainen 1999).

The ‘full’ and ‘reduced’ tagsets used with Morphy refer to the way tagging errors were counted; with the ‘full’ tagset, the whole morpho-syntactic description had to be correct, i.e., if any part of it was incorrect – e.g., if the case was given as ‘dative’ instead of ‘nominative’ (a fairly common error in our texts) – the error count would be increased by 1. In the case of the ‘reduced’ set, however, a correct part of speech[7], together with an error or errors in gender, case, and number for nominal parts of speech, and person/number for finite verbs, only would count as 0.25 errors.

The results seem to show that tagger performance is dependent to some extent on text type, but at the present time we can only note this as a topic which merits further investigation.

2.2.3 Step 3: Finding tagger differences

Next, a correspondence table was constructed for the tagsets of the taggers, and a tagger comparison program (described by Borin *et al.* Forthcoming) was used on their output. The hypotheses to be tested were:

(1) there would be differences between the two taggers in the errors made

(2) these differences would show some systematicity, which could be utilised to improve tagging accuracy by combining the two taggers.

Both hypotheses were supported by the results of the experiment. There were differences between the taggers (see Table 2), and some of the differences turned out to be systematic.

Table 2: Tagger differences: Which tagger was right how often?

Corpus	*Morphy*	*TreeTagger*	*Neither*	*Total*
SGP	101 / 35.5%	176 / 62%	7 / 2.5%	284 / 100%
Scania	86 / 36.1%	139 / 58.4%	13 / 5.5%	238 / 100%
Total	187 / 35.8%	315 / 60.4%	20 / 3.8%	522 / 100%

[7] Here we used, roughly, the part-of-speech inventory of TreeTagger, so that, e.g., finite verbs, infinitives, and participles were counted as different parts of speech, even though they have the common major part of speech “VER” in Morphy's tag set.

2.2.4 Step 4: Finding the systematic differences

Finding the systematic differences between POS taggers implies making a decision as to which variables should be taken into account, i.e. should provide the input parameters for the if–then rules which should be the result of the next step. This amounts to a hypothesis about which factors influence tagging performance, and our initial hypothesis has been that the following parameters would be relevant:

- the individual tags themselves;
- disjunctions of tags, denoting linguistically natural categories, e.g., both common nouns and proper nouns are nouns, both verbs and adjectives are verbal words in many languages;
- the text type, in our case the technical text of the Scania corpus vs. the administrative-political text type of the SGP;

2.2.5 Step 5: Formulating rules for combining taggers

Using the differences between taggers and the hypothesis about which parameters were likely to influence tagger performance, rules were formulated to choose the output of the inferior tagger (Morphy) over that of the better tagger (TreeTagger) under certain, systematically recurring conditions. The general format of the rules is:

if Morphy and TreeTagger assign non-equivalent tags to a text word,
and the following conditions (see Table 3) are fulfilled,
then choose the tag that Morphy assigned,
else choose the tag that TreeTagger assigned.

Table 3 shows the conditions inferred from a linguistic analysis of tagging differences. Tendencies as well as absolute conditions were taken into account, and the last two columns in Table 3 show how often (in the examined material) a rule using the current condition would pick a correct tag ("+ cases" in Table 3), and how often it would be wrong (the "– cases").

Table 3: Conditions for choosing Morphy (then-clause in rule)

Text type	*TreeTagger tag(s)*	*Morphy tag(s)*	*+ cases*	*– cases*
both	–	ABK	33	0
Scania	ADJA, ADJD	SUB *, EIG *	15	1
both	ADJD	VER PA2	7	2
SGP	ADJD	VER *	4	0
SGP	ADV	KON *	13	0
both	NN	ADJ (–ADV)	15	3

When the rules were applied to the TreeTagger evaluation sentences (see above), the accuracy figures shown in Table 4 were obtained (the previous results, from Table 1, are repeated here for comparison).

Table 4: Accuracy of combined taggers

Tagging regime	*Scania*	SGP
TreeTagger only	96.3%	96.2%
TreeTagger + Morphy	96.7%	97.8%
Difference (% units)	+0.4	+1.6

We see that there was an improvement in tagging performance, even a marked improvement in the case of the SGP texts. We must remember that, here, an improvement of even a single percent unit is much, considering that the span between the chance baseline and maximum human interjudge agreement is less than 10 percent units (Voutilainen 1999).

2.3 Combining word alignment and tagging

Is it a reasonable assumption, as made, e.g., by Melamed (1995) "that word pairs that are good translations of each other are likely to be the same parts of speech in their respective languages"?

From a purely linguistic standpoint, there is reason to doubt that this assumption holds for the general case of any language compared with any other language, and for any part of speech. It has been held for a long time in linguistics that nouns and verbs are the only universal parts of speech, in the sense that they are found in all human

languages (and it seems that even verbs are not all that necessary; cf. Pawley 1993).

Even though not universally valid, one might entertain the hypothesis that the assumption is more likely to hold for languages which either are closely related genetically – like Swedish and German – or have been in contact for a long time – as in the case of Swedish and Finnish.

Even in the latter case, it is conceivable that not all parts of speech are equally likely to remain invariant when translating from one language to the other. If we could determine under what circumstances this is likely to be the case, we would be in possession of a very useful piece of knowledge, since we could then (partly) replace, as it were, tagging of L2 by the alignment with a tagged parallel text in L1. In the case where we do not have a tagger for L2, but one for L1, and, in addition, a parallel L1 text to the L2 text that we would like to tag, we could then utilise this knowledge to tag the L2 text with the help of the L1 tagger and a word alignment algorithm.

In order to test these hypotheses, one should test them with many language pairs, plotting the result against the degree of relatedness among the languages and the various parts of speech. Here, we make a start in this direction by investigating the language pair Swedish and German. The investigation proceeded as follows.

(1) A Swedish–German parallel text was word aligned with a word alignment tool developed in our department (Tiedemann forthcoming). The text was one the SGP texts in the ETAP corpus. The word alignment recall was slightly below 40%, i.e. 40% of the potential word alignments in the test sentences were actually returned by the word alignment program. Some of the alignments found are shown in Table 5;

(2) The German text was POS tagged with Morphy (because of the larger tag set);

(3) For every German word–tag combination, if there was a word alignment with a Swedish word, that word was assigned the POS tag of the German word;

(4) The accuracy of the POS tags assigned in the previous step was assessed manually, using a version of the SUC tag set.

German and Swedish POS tags are not directly comparable, of course. Thus, we decided to look primarily at major part-of-speech correspondences, but with an eye to possible subcategory correspondences.

Table 5: Some Swedish–German word alignments

svdeprf83:[8]	
Industrins **NN SIN**	Industrie **Industrie SUB GEN SIN FEM**
anpassning **NN SIN**	Anpassung **Anpassung SUB NOM SIN FEM**
krav **NN *SIN/PLU**	Anforderungen **Anforderung SUB AKK PLU FEM**
och **KN**	und **und KON NEB**
processer **NN PLU**	Prozesse **Prozeß SUB NOM PLU MAS**
produkter **NN PLU**	Produkte **Produkt SUB DAT SIN NEU**
renare **JJ**	reiner **rein ADJ ADV**
skall **VB**	sollen **sollen VER MOD 3 PLU**
svdeprf102:[9]	
Livsmedelskontrollen	Nahrungsmittelkontrolle
NN SIN	**Nahrungsmittelkontrolle SUB NOM SIN FEM**
skärps ***VB**	verschärft **verschärfen VER PA2**

In Table 6, major POS tag correspondences for the aligned units are shown, i.e. we see how many of the German POS tags would also have been appropriate POS tags for the Swedish words with which they are aligned. The question is posed both for complete POS tags, but also for the main category (VER, SUB, etc) part of the Morphy tags.

Table 6: Main POS category (VER [V], SUB [N], etc.) correspondences in aligned words, arranged by correct and incorrect word alignments.

Correct alignments (64 of 78)		Incorrect alignments (14 of 78)	
correct POS	incorrect POS	correct POS	incorrect POS
61 (95%)	3 (5%)	1 (8%)	13 (92%)

[8] Sentence alignment unit **83** in the Swedish [**sv**] – German [**de**] parallel SGP [=Sw. **RF**] corpus.

[9] Sentence alignment unit **102** in the Swedish [**sv**] – German [**de**] parallel SGP [=Sw. **RF**] corpus.

It turned out that Morphy POS tag subcategories (i.e., inflectional information) were, in general, not relevant, with one exception: For the NN (Morphy: SUB) subcategory 'number' (7 PLU, 22 SIN in the text), the German value turned out to be the correct choice for the Swedish correspondence 27 times out of 29.

We also see that correct alignments and correct POS tags go hand in hand (as shown in the lower left part of Table 6), while bad alignments also imply bad POS assignments (the lower right part of the table).

In brief, the conclusion tentatively to be drawn from this experiment is that the idea of using word alignment as a stand-in for, as it were, or as a complement to, POS tagging is viable and worth exploring further. However, it seems that certain prerequisites have to be fulfilled for it to work:

- The languages in question should be genetically or typologically close;
- A high word alignment precision is needed;
- Only coarse-grained POS tagging – i.e., on the level of the main syntactic category, but not with regard to finer morphosyntactic distinctions – seems possible with this approach.

3 Conclusions and future work

I have tried to show you two examples of how existing knowledge sources can be brought together in novel ways to solve a particular task, that of POS tagging a multilingual parallel corpus. It turns out that they jointly – much like a team of cooperating humans – will achieve a better result than any single one of them – any single individual, in the human analogy – could achieve on its own.

Of course, there is an additional knowledge source involved here, namely the linguistic knowledge of the investigator, since the outcome of both approaches outlined above is a set of rules formulated on the basis of that knowledge, used to extract linguistically relevant generalisations from the results of the experiments, and not on the basis of, e.g., a statistical model.

There are many directions in which this research could be continued. In particular, we can discern at least the following strands of

inquiry, which all are worth pursuing, individually or in various combinations:

- trying to clarify the roles of tagger technology, text type, training corpus size, tag set size, etc., i.e. all the variables that presumably play a role in determining tagger performance, in order to make more informed decisions as to if and how POS taggers are to be combined in order to enhance their performance;
- investigating whether the first procedure outlined above could be extended to the partial mistaggings made by a tagger like Morphy. If this turned out to be the case (although I suspect that it will not), one could use the larger (hence more fine-grained) tag set of Morphy with the greater precision of TreeTagger;
- exploring machine learning methods as a way to automatise the rule formulation step in this procedure. The question is which method(s) to investigate, but a natural first candidate would be transformation-based learning (TBL), as we have some experience of working with this method in the project (Prütz forthcoming);
- extending the second kind of investigation presented above to other languages and language pairs, in order to tease out the relevance of such factors as the influence of typological and genetic parameters on the results.

References

Borin, Lars, Camilla Bengtsson and Henrik Oxhammar. Forthcoming. Comparing and combining part of speech taggers for multilingual parallel corpora. Research report etap-rr-03. Dept. of Linguistics, Uppsala University.

Brill, Eric and Jun Wu. 1998. Classifier combination for improved lexical disambiguation. *Proceedings of COLING-ACL'98*. Montreal: Université de Montréal.

Cunningham, H., R.G. Gaizauskas and Y. Wilks. 1995. A general architecture for text engineering (GATE) – a new approach to language engineering R&D. Technical Report CS-95-21. Department of Computer Science, University of Sheffield.

Ejerhed, Eva and Gunnel Källgren. 1997. Stockholm Umeå Corpus Version 1.0, SUC 1.0. Department of Linguistics, Umeå University.

Godfrey, John J. and Antonio Zampolli. 1997. Overview. *Survey of the State of the Art in Human Language Technology*, ed. by Ronald Cole et al. Cambridge: Cambridge University Press. 381–384.

van Halteren, Hans, Jakub Zavrel and Walter Daelemans. 1998. Improving data driven wordclass tagging by system combination. *Proceedings of COLING-ACL'98*. Montreal: Université de Montréal.

Lezius, Wolfgang, Reinhard Rapp and Manfred Wettler. 1998. A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. *Proceedings of COLING-ACL'98*. Montreal: Université de Montréal.

Màrquez, Lluís, Lluís Padró and Horacio Rodríguez. 1998. Improving tagging accuracy by using voting taggers. *Proceedings of NLP+IA/TAL+AI'98.* Moncton, New Brunswick, Canada.

Mason, O. 1997. QTAG – A portable probabilistic tagger. Corpus Research, University of Birmingham.

Melamed, Dan. 1995. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. *Proceedings of the Third Workshop on Very Large Corpora*. Boston, Massachusetts.

Pawley, Andrew. 1993. A language which defies description by ordinary means. *The Role of Theory in Language Description*, ed. by William A. Foley. Berlin: Mouton de Gruyter. 87–129.

Prütz, Klas. Forthcoming. Part-of-speech tagging for Swedish. *Parallel Corpora, Parallel Worlds*, ed. by Lars Borin. Dept. of Linguistics, Uppsala University.

Schiller, Anne, Simone Teufel, Christine Stöckert and Christine Thielen. 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Draft. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung / Universität Tübingen, Seminar für Sprachwissenschaft.

Tiedemann, Jörg. Forthcoming. Word alignment step by step. *Proceedings of the 12th Nordic Conference on Computational Linguistics (Nodalida99),* ed. by Torbjørn Nordgård. Trondheim.

Tumer, Kagan and Joydeep Ghosh. 1999. Linear and order statistics combiners for pattern classification. *Combining Artificial Neural Networks*, ed. by Amanda Sharkey. Berlin: Springer-Verlag. 127–162.

Voutilainen, Atro. 1999. An experiment on the upper bound of interjudge agreement: The case of tagging. *Proceedings of EACL'99.* Bergen, Norway: University of Bergen. 204–208.