Linguistic Indeterminacy as a Source of Errors in Tagging

Gunnel Källgren
Department of Linguistics
Stockholm University
S-106 91 Stockholm
Sweden
gunnel@ling.su.se

Abstract

Most evaluations of part-of-speech tagging compare the utput of an automatic tagger to some established standard, define the differences as tagging errors and try to remedy them by, e.g., more training of the tagger. The present article is based on a manual analysis of a large number of tagging errors. Some clear patterns among the errors can be discerned, and the sources of the errors as well as possible alternative methods of remedy are presented and discussed. In particular are the problems with undecidable cases treated.

1 Background

When the performance of automatic part-of-speech taggers is discussed, it is normally measured relative to some standard material, such as the Brown Corpus, or to a manual tagging or a manual proof-reading of (some smaller part of) the tagged material. The performance of the automatic tagger is calculated as the difference between the standard material and the output of the tagger to be evaluated, with all differences regarded as errors by the tagger.

In a study carried out on material from a large Swedish corpus, Källgren (1996) made a careful inspection of all instances where a manual and an automatic tagging differed in a material of 50,000 words of balanced text. The differences were classified as 'man errors', 'machine errors' or errors common to both. The errors were furthermore classified according to type, and some clear patterns could be seen. The present article picks up some of the findings and looks closer at a kind of error which Källgren calls 'mirror image errors', where two readings of a word are constantly mixed up with each other in both directions. Errors of this kind have been noted by others as well, and solutions to the problems they cause have been suggested. Some such suggestions and the possible outcome of their application to the Swedish material will be discussed in the following.

2 The Linguistic Material Used in the Study

The error analysis on which this study is based was carried out on material from the Stockholm-Umeå Corpus of modern written Swedish. (See Källgren 1990.) It is a carefully composed, balanced corpus. Its composition follows the principles established by the Brown and LOB corpora, with adjustments for the fact that it should cover the most common genres of the Swedish of the 1990's. It contains newspaper texts, fact, and fiction on several stylistic levels. The texts all consist of written prose published sometime between 1990 and 1994. No spoken language material is included in the corpus.

All words in the SUC are tagged for part-of-speech and for inflectional features. For a description of the SUC annotation system, see Ejerhed et al. (1992). The tagged texts of the SUC are converted into SGML format and additional tags are added in accordance with the TEI Guidelines (Sperberg-McQueen and Burnard 1993, Källgren 1995) to give the format in which the corpus will finally be distributed. There are legal permissions allowing the corpus to be used and distributed for non-commercial research purposes.

3 Manual and Automatic Markup

The SUC has been annotated by a process that combines automatic and manual steps. The raw texts get their first analysis from the SWETWOL computerized dictionary (Karlsson 1992) and then pass a step of postprocessing to reach the analysis described in the SUC tagging manual (Ejerhed et al. 1992). The coverage of the dictionary is high, but the degree of ambiguity in Swedish is also high, actually higher than in English, so the texts return from dictionary lookup with 51% of the word tokens carrying more than one analysis.

In the next step, a human annotator is to mark for each ambiguous word which of the suggested readings is the correct one and for each unambiguous word whether the suggested reading is correct. The output of this step is used as the 'man version' in the

man-machine comparison (or rather the 'woman version' as the majority of the annotators were female students).

The entire corpus of 1 million words has passed through this stage of manual disambiguation and annotation, which makes it an important standard that can be used as a tool, e.g., when training probabilistic taggers. The goal of the experiment reported in Källgren (1996) was, however, to compare 'sheer' machine tagging to the performance of human annotators. The tagger used is thus one that does not need tagged and disambiguated material to be trained on, namely the XPOST originally constructed at Xerox Parc (Cutting et al. 1992, Cutting and Pedersen 1993).

The XPOST algorithm has been transferred to other languages than English. Douglass Cutting himself made the first Swedish version of it (Cutting 1993) and a later version has been implemented by Gunnar Eriksson (Eriksson 1995) and refined by Tomas Svensson (Svensson 1996). It is this latter version that has been used in the experiment.

Starting from a set of texts and a lexicon, the XPOST looks up all words in the texts and assigns to them a set of one or more readings. The words are then classified into so-called ambiguity classes according to which set of readings they have been assigned. The training is performed on ambiguity classes and not on individual word tokens. Källgren (1996) gives a more covering description of how XPOST is used on the Swedish material and also sketches the major differences between this algorithm and some others used for tagging, such as PARTS (Church 1988) and VOLSUNGA (DeRose 1988).

A characteristic feature of the SUC is its high number of different tags. The number of part-of-speech tags used in the SUC is 21. With the addition of a category for foreign words the number of major categories used is 22 (plus three tags for punctuation), which is in no way a remarkable amount, but the SUC tags are composite. This means that all words have one tag for part-of-speech, but for many parts-of-speech this tag is followed by other tags for various morphological features. Where, e.g., English nouns have a variation between two possible values, singular and plural, the Swedish pattern allows for 1 x 2 x 2 x 2 x 3 = 24 different tags, specifying not only part-of-speech but also gender, number, definiteness, and case. The number of different tags actually occurring in texts is mostly around 180.

A remarkable fact is that the high number of different tags does not seem to influence the training and performance of probabilistic taggers negatively in the way that might have been expected. The morphological errors in the material are not disturbingly many, considering the fact that all Swedish content words have such features. Morphological agreement provides enough information to make it possible for an automatic tagger to pick the right form in most cases. This sensitivity to close context probably explains why the high number of tags does not influence performance when it comes to picking an alternative, but it does not explain why training is so little affected by the high number of different observed situations.

4 Results from a Comparison between 'Man' and 'Machine'

The automatic tagger was run on 50,000 words of text not used in the training of the tagger. The output was compared to the same texts with manual disambiguation. All instances where the two differ have been manually inspected. The evaluation of the results is far from trivial. The 'correctness' of the tagging must be judged relative to some norm. One such norm is the SUC tagging manual (Ejerhed et al. 1992). Although it is very comprehensive and explicit, no manual can ever foresee and cover all the tricky instances that will occur in unrestricted language. Another norm is the intuition of the working linguist, with the possibility of consulting other people to get their intuitions. This also has clear drawbacks. There will always remain a set of doubtful cases which do not necessarily depend on deficits in the linguistic description. Be it here sufficient to say that in general I prefer the term 'consistent (with a certain norm)' instead of the term 'correct'; nevertheless, in the following discussion I will call the deviances from the applied norm 'errors'.

Table 1 gives the errors found in a material of 50,498 words sorted according to whether they occurred in automatically or manually tagged text or both. Where both have an error, the errors can sometimes be of the same type, sometimes of different types.

Table 1. Tagging Errors According to Source

	N	%
Errors only in automatic tagging	3591	7.1
Errors only in manual tagging	503	1.0
Errors in both	110	0.2
Total	4204	8.3

The automatic tagger is truly automatic in that it has not at all been adjusted to the specific task at hand. With fairly little trimming it could well reach a level of at least 95-96% consistence with the human annotator but now the basic idea was to test it 'raw'. Humans are not infallible, if anyone thought so, 1.2% of the errors are man-made. It is still a consolation to see that human annotators are seven times as good as computers when it comes to disambiguation.

5 Types of Errors

The errors occurring in the material can be classified according to type. By 'error type' is here meant a classification of tag pairs with an erroneous tag followed by the correct tag, e.g., an error can be of the type 'preposition suggested where it should have been an adverb'. This classification shows both which parts-of-speech are most often involved in errors and which readings of a particular word are most often mixed up with each other, and in which direction the errors mostly go. The classification can also give hints about what could possibly be done about the errors.

5.1 Errors among Content Words

It is clear that content words (here: nouns, verbs, adjectives, participles, proper nouns) are seldom involved in errors. Considering the large proportion of the number of running words that these major categories cover, this is even more remarkable. If words from these categories are ever mixed up, they are mixed up in very specific patterns, namely with themselves (as when different inflected forms of the same stem coincide) or they are mixed up with words they are related to (e.g., by derivation). Among the ten most common error types for either automatic or manual disambiguation, there are actually only two that involve content words.

One of these error types is almost exclusively in the realm of automatic disambiguation. Swedish nouns are inflected according to five different declensions, one of which has zero plural. The automatic tagger sometimes mistakes singular nouns of that declension without modifiers for plurals, but never the other way round. This is just as could be expected; 'naked' plurals are far more common than 'naked' singulars in all declinations and will thus be favoured by the statistics. To remedy this situation, it would probably be necessary to have a phrasal lexicon, as most instances of naked singular nouns appear in lexicalized phrases.

As has been pointed out for English material (cf. below) different inflections of the same verb can get mixed up. This phenomenon can be found in Swedish too, but not very frequently.

The other common error type involving content words concerns adverbs derived from adjectives. The most frequent derivational pattern for Swedish adverbs makes them identical to neutral singular indefinite adjectives. Here both manual and automatic disambiguation leads to errors but in different directions. The automatic tagger suggests adverb where there should have been an adjective, while human annotators sometimes call an adverb an adjective. Both types mainly occur post-verbally and often at the very end of a graphic sentence, where it may be difficult to decide whether the concerned word is a predicative adjective or an adverb. It may well be that a subcategorization of verbs might eliminate the problem, but this is a large task to implement both in the lexicon and in the tagger.

However, these errors are neither the most frequent nor the most disturbing ones. Instead, it is the function words that get mixed up in all their different uses. Actually, almost all errors concern function words and a scrutiny of them makes it clear how doubtful the whole concept of correctness is in this connection.

5.2 Errors among Function Words

The degree of homography - or is it polysemy? - is generally higher among function words than among content words which, of course, leads to more situations where errors can occur. Furthermore, the number of readings connected with each word token is highly dependent on the linguistic description used as a basis for the tagging system, its theoretical assumptions and the granularity of the system, among other things.

The ten words most frequently involved in errors in the studied material are (with approximate translations and number of errors in parenthesis) the following: 'det' (*it/the* in neuter gender, 330 errors), 'ett' (*a/one* in neuter, 254), 'som' (rel.pron and adv., 180), 'den' (*it/the* in common gender, 153), 'om' (*if, about*, 122), 'en' (*a/one* in common, 109), 'att' (*that*, inf.marker, 83), 'så' (*so*, 79), 'ut' (*out*, 73), 'för' (*for*, 70). They are all high frequency function words that play many different syntactic roles depending on their context.

One interesting fact that the classification into error types makes clear is that all the different readings of these words do not get mixed up at random but in rather strong, often mirror-like patterns. Let us take the word 'om' as an example. It can be used as adverb, preposition, or subordinating conjunction and all the six possible mistagged combinations do occur, but with quite varying frequency. Three of them are almost neglectable and one has a strong unidirectional pattern where the reading as an adverb (more precisely a verbal particle) is often taken for a preposition. This is an instance of the by far most common error type in the entire material, and is of course directly dependent on the way verbal particles are treated in the underlying linguistic description.

The remaining two error types are the most interesting ones. They form a bidirectional pattern where the reading as a preposition is confused with the reading as a subordinating conjunction. Preposition instead of subjunction appears 40 times, subjunction instead of preposition 33 times, altogether 77 of the 122 errors connected with the word 'om'. All errors on this word were machine-induced, except 8 cases where human annotators took a subjunction to

be a preposition. Some of the error situations may be regarded as truly undecidable.

6 Tagging Undecidable Situations

How are bidirectional error patterns like the one above to be treated? Looking at their close context, it is often impossible to handle the situation with some smart tagging restriction or other device. They are also so equal in number and so frequent that one cannot simply decide to let one reading overrule the other and live with the errors that such a happy-go-lucky solution would give rise to. (As a practicing corpus tagger, I know that this unorthodox method can sometimes be the best way out of problematic situations.)

Another possibility would be to amalgamate the two readings into one, bivalued or underspecified, depending on how one chooses to see it. As already mentioned, these more or less undecidable bidirectional patterns have been observed and discussed by others working with the tagging of large corpora, and they have, seemingly independently of each other, come up with similar suggestions. Below are three quotations dealing with this matter.

The Penn Treebank: 'However, even given explicit criteria for assigning POS tags to potentially ambiguous words, it is not always possible to assign a unique tag to a word with confidence. Since a major concern of the Treebank is to avoid requiring annotators to make arbitrary decisions, we allow words to be associated with more than one POS tag. Such multiple tagging indicates either that the word's part of speech simply cannot be decided or that the annotator is unsure which of the alternative tags is the correct one.' (Marcus et al. 1993, 316.)

The British National Corpus: 'In order to provide more useful results in a substantial proportion of the residual words which cannot be successfully tagged, we have introduced portmanteau tags. A portmanteau tag is used in a situation where there is insufficient evidence for Claws to make a clear distinction between two tags. Thus, in the notoriously difficult choice between a past participle and the past tense of a verb, if there is insufficient probabilistic evidence to choose between the two Claws marks the word as VVN-VVD. A set of fifteen such portmanteau tags have been declared, covering the major pairs of confusable tags.' (Garside 1995.)

Constraint Grammar: 'In the rare cases where two analyses were regarded as equally legitimate, both could be marked.' (Voutilainen and Järvinen 1995, 212.)

It is, however, important that the situations where underspecified tags can be used are restricted to well-defined cases and that the reasons for using them are quite clear. They should have what I call a 'mirror' character, in that the interchange goes in both directions, and they should concern clearly distinct pairs of tags even when a word has several other tags as well. Such situations are more common in automatic tagging but they occur in manual tagging as well.

The reasons for a situation being undecidable can, however, vary. Voutilainen and Järvinen, in their study of inter-annotator agreement, mention three situations where an underdetermined analysis was accepted:

'When the judges disagree about the correct analysis even after negotiations. In this case, comments were added to distinguish it from the other two types. Neutralisation: both analyses were regarded as equivalent. (This often indicates a redundancy in the lexicon.) Global ambiguity: the sentence was agreed to be globally ambiguous.' (Voutilainen and Järvinen 1995, 212.)

Marcus et al. (1993) allow underspecified tagging both for annotators' uncertainty or disagreement and for cases that correspond to Voutilainen and Järvinen's neutralisation and global ambiguity. This may be infelicitous. It is important to keep a clear borderline between situations that could be solved in principle and such that are truly undecidable. The latter ones may lead us to questions about the nature of language and to what extent natural language really is exact and welldefined.

Introducing underspecified tags would influence the training and performance of a probabilistic tagger in at least the following ways: a) The concerned words would mostly get more alternative tags, one for each of the unambigous readings plus one for the underspecified one. According to common tagging principles, this would be a disadvantage. b) There would be fewer observations of each of the alternative tags, as the competing unambiguous tags both would lose some of their instances to their common underspecified alternative. This would also be a disadvantage. c) The observations of each tag would hopefully be more correct, as the instances 'lost' to the underspecified tag would be the tricky and atypical cases that otherwise might obscure the contextual patterns of the unambiguous tags. d) The underspecified instances can later be automatically retrieved for either manual inspection or some more elaborate disambiguation device.

It is still an open question whether the more clear-cut distinctions introduced by the underspecified tags compensate for the accompanying disadvantages, but at least they have the intellectually pleasing property of showing where there are truly ambiguous situations in language. By systematic modifications of the tagset along these lines it is possible to decide to what extent the introduction of underspecified tags will improve the overall performance of a tagger and/or facilitate the task of human annotators.

References

Church, K. W. (1988), 'A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text', in Proceedings of the Second Conference on Applied Natural Language Processing (ACL), 136-43 (Austin).

Cutting, D. (1993), 'Porting a Stochastic Part-of-Speech Tagger to Swedish', in Eklund, R. (ed.) Nordiska Datalingvistikdagarna, 65-70 (Stockholm).

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992), 'A Practical Part-of-Speech Tagger', in Proceedings of the Third Conference on Applied Natural Language Processing (ACL) (Trento).

Cutting, D. and Pedersen, J. (1993), The Xerox Part-of-Speech Tagger, Xerox PARC technical report.

DeRose, S. J. (1988), 'Grammatical Category Disambiguation by Statistical Optimization', Computational Linguistics, Volume 14:1, 31-9.

Ejerhed, E., Källgren, G., Wennstedt, O. and Åström, M. (1992), The Linguistic Annotation System of the Stockholm-Umeå Corpus Project, version 4.31. Publications from the Department of General Linguistics, University of Umeå, no. 33.

Eriksson, G. (1995), 'Beskrivning av arbetet med att utveckla en XPOSTtagger för svenska', Technical report, Telia Research Infovox (Stockholm).

Garside, R. (1995), 'Using CLAWS to Annotate the British National Corpus', URL: http://info.ox.ac.uk/bnc/garside_allc.html.

Källgren, G. (1990), '"The First Million is Hardest to Get": Building a Large Tagged Corpus as Automatically as Possible', in Proceedings from Coling '90 (Helsinki).

Källgren, G. (1995), 'Manual for TEI conformant mark-up of the SUC', draft version, Department of Linguistics, Stockholm University.

Källgren, G. (1996), 'Man vs. Machine - Which is the Most Reliable Annotator?', to appear in Perissinotto, Giorgio (ed.), Research in Humanities Computing 6, Oxford University Press.

Karlsson, F. (1992), 'SWETWOL: A Comprehensive Morphological Analyzer for Swedish', in Nordic Journal of Linguistics Vol. 15:1-45.

Marcus, M. P., Marcinkiewicz, M. and Santorini, B. (1993), 'Building a Large Annotated Corpus of English: The Penn Treebank', in Computational Linguistics Volume 19:2, 313-30.

Sperberg-McQueen, C. M. and Burnard, L. (1993) (eds.), Guidelines for Electronic Encoding and Interchange (Chicago, Oxford).

Svensson, T. (1995), 'Om tagguppsättningar i en första ordningens gömd Markovmodell', Technical report, Telia Research Infovox (Stockholm).

Voutilainen, A. and Järvinen, T. (1995), 'Specifying a shallow grammatical representation for parsing purposes', in Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, 210-14.