Morphological Tagging Based Entirely on Bayesian Inference

Christer Samuelsson
Stockholm

Abstract

The influence of various information sources on the ability of a statistical tagger to assign lexical categories to unknown words is investigated. The literal word form is found to be very much more important than other information sources such as the local syntactic context. Different ways of combining information sources are discussed. Methods for improving estimates based on scarce data are proposed and examined experimentally.

1 Introduction

Tagging is the art of assigning a specific label, a tag, to each word in a corpus from a pre-defined set of labels — a (tag) palette. The traditional problem addressed is that of disambiguation, see for example Church (1988). A lexicon states what different tags each word can possibly be assigned to, and for any specific word, this is a small subset of the palette.

Normally, the most likely assignment of tags to the words of a corpus is determined by statistical optimization using dynamic programming techniques, as is well-described in for example DeRose (1988). Some existing taggers, though, make use of hand-coded heuristic rules to guide the assignments, see for example Brill (1992) and Källgren (1991). Until recently, empirical results have indicated that statistical methods are superior to rule-based ones. However, the results reported in Voutilainen *et al* (1992) indicate that this may actully not be the case. However, rule-based approaches suffer from the major disadvantage of being *very* labour intense.

The approach taken here differs somewhat from mainstream tagging endeavours. Our main goal is not disambiguation, but to investigate the influence of various information sources, and ways of combining them, on the ability to assign lexical categories to unknown words. Here, we dispense with heuristic rules and lexical entries altogether, and instead rely entirely on the power of statistics to extract the required information from a pre-tagged corpus during a training phase. This has the advantage of making the tagger completely language independent.

The information sources employed are the literal appearance of the word and the tags assigned to the neighbouring words. The way these sources

Proceedings of NODALIDA 1993, pages 225-238

of information are combined is novel for tagging applications. Another important innovation is the "successive abstraction" scheme for handling scarce training data by generalizing to successively wider contexts. A final original feature is the fact that the tagger is implemented entirely in SICStus Prolog.

Bayesian inference is used to find the tag assignment T with highest probability $P(T\ M,S)$ given morphology M (word form) and syntactic context S (neighbouring tags). This quantity is calculated from the probabilities $P(T\ /M)$ and $P(T\ /S)$. Before describing how the latter two quantities are estimated, we will address two important issues, namely those of combining them and of estimating the probabilities of events for which there is only a small number of observations.

2 Combining information

Several methods seem currently to be in use for combining information sources. One method is to simply set the probability of the hypothesis H given the combined evidence to the product of the probabilities of the event given each context:

$$P(H \mid M,S) \approx P(H \mid M) \bullet P(H \mid S)$$

Unfortunately, this can lead us far astray from the correct figure. For example, using Bayes' inversion formula, and assuming that these information sources are independent, and conditionally independent given the hypothesis H, i.e.

(1)

$$P(M,S) = P(M) \bullet P(S)$$
$$P(M,S \mid H) = P(M \mid H) \bullet P(S \mid H)$$

will yield us the exact formula

$$P(H \mid M,S) = (P(H \mid M) \bullet P(H \mid S)) / P(H)$$

This means that unlikely hypotheses will be unduly penalized in the above approximation by omitting the denominator $P(H)$. Even if these assumptions are not valid, this line of reasoning shows that intuitively, an extra factor proportional to the probability $P(H)$ is introduced.

Another method is to use a weighted sum of the probabilities:

$$P(H \mid M,S) \approx \lambda_M P(H \mid M) + \lambda_S P(H \mid S)$$

In general, the weight assigned to the morphological probability (λ_M) will be much larger than that assigned to the syntactic probability. The problem with this approach is that these weights are static, and not dependent on the relative predictive power of the two information sources in each particular case.

The approach taken here remedies these shortcomings: We wish to estimate the probability $P(H \mid e_1,\ldots,e_n)$ of the hypothesis H given the evidence e_i: $i = 1,\ldots,n$. We will go by way of the posterior odds $O(H \mid e_1,\ldots,e_n)$ (see for example Pearl (1988), pp. 34–39), which are defined by

$$O(A \mid B) = P(A \mid B) / P(\neg A \mid B) = P(A \mid B) / (1 - P(A \mid B))$$

We will make the independence assumption

(2)
$$O(H \mid e_1,\ldots,e_n) / O(H) \approx (O(H \mid e_1) / O(H)) \bullet \ldots \bullet (O(H \mid e_n) / O(H))$$

In our case the hypothesis H is the tag T and the evidence e_1 and e_2 are the word form M and the syntactic context S. Thus:

$$O(T \mid M,S) \approx (O(T \mid M) \bullet O(T \mid S)) / O(T)$$
$$P(T \mid M,S) = O(T \mid M,S) / (1 + O(T \mid M,S))$$

This formula has several advantages. Firstly, it is exact under the independence assumptions of Eq.(1), with the additional assumption

(3)
$$P(M,S \mid \neg H) = P(M \mid \neg H) \bullet P(S \mid \neg H)$$

which proves that it doesn't introduce an extra factor of $O(H)$. The relationship between equations Eq.(1), Eq.(3) and Eq.(2) is that the two former together imply the latter, but not vice versa. Secondly, using the odds instead of the probabilities has a stabilizing effect when none of the independence assumptions are valid. Thirdly, the impact of each of the two sources of information is allowed to depend dynamically on how much distinctive power they carry, rather than being prescribed beforehand, as is the case when using a weighted sum.

3 Handling scarce data

We now turn to the problem of estimating the statistical parameters for which there is only a small amount of training data.

3.1 Successive abstraction

Assume that we want to estimate the probability $P(E \mid C)$ of the event E given a context C from the number of times N_E it occurs in N trials, but that this data is scarce. Assume further that there is abundant data in a more general context $C' \supset C$ that we want to use to get a better estimate of $P(E \mid C)$.

If there is an obvious linear order $C' = C_m \supset C_{m-1} \supset \ldots \supset C_1 = C$ of the various generalizations C_k of C, we abstract successively to the lowest k for which data suffices. We will refer to this as "linear abstraction". A simple example is estimating the probability $P(T \mid l_n, l_{n-1}, \ldots, l_{n-j})$ of a tag T given the last $j+1$ letters of the word. The estimate will be based on the estimates of $P(T \mid l_n, l_{n-1}, \ldots, l_{n-j})$, $P(T \mid l_n, l_{n-1}, \ldots, l\_\{n-j+1\})$, ... ,$P(T \mid l_n, l_{n-1}, \ldots, l_{n-k})$, where k is the smallest number for which there is a good estimate of $P(T \mid l_n, l_{n-1}, \ldots, l_{n-k})$.

Even if there is no obvious linear order of the various generalizations, they might stem from a small number of sources, each of which has a linear order. An example is generalizing compound nouns using a sortal hierarchy. Here, the possible generalizations of the compound are the compounds of the generalizations of each noun. This can be used to explore the possible generalizations systematically and facilitates pruning using a heuristic quality measure of the estimates. If this measure is simply the total number of observations in the next generalized context, we will call this "greedy abstraction".

3.2 Improving estimates

Several different methods were tried for combining the estimates based on scarce data with estimates from a more general context — in Section 3.2.1 a confidence interval is used and in Section 3.2.2 weighted sums.

3.2.1 Using a confidence interval

To calculate an estimate $\boldsymbol{p}$ of the probability $p = P(E \mid C)$ we will use the fact that the quotient $\xi_n = N_E / N$, where N_E is the number of times E occurs in N trials, is a random variable with a binomial distribution, i.e.

$$\xi_n = N_E / N \sim \text{bin}(p, \sqrt{(p(1-p) / N)})$$

to get a first estimate $\boldsymbol{x}$ of p and a confidence interval $x_1 < p < x_2$ with confidence degree α, where $x_1 < \boldsymbol{x} < x_2$. Given these quantities,

- **If** for a pre-defined threshold θ,
  \label{EqThres}
  $$(x - x_1) / x < \theta \text{ and}$$
  $$(x_2 - x) / x < \theta$$

- **Then** set $p = x$. Here, we are confident ($100 \cdot \alpha$ %) that x is a good ($\pm 100 \cdot \theta$ %) estimate, and are satisfied with this.

- **Else** generalize the context C one step to C'; calculate an estimate $\hat{p}'$ of the probability $P(E \mid C')$ recursively; let $f(x)$ be some suitable function and set $p = f(\hat{p}')$. Examples of such functions are discussed below. Thus, $f(x)$ is used to let the observations of E in context C guide the estimate according to their reliability.

- Re-normalize so that $P(\Omega \mid C) = 1$ for the entire sample space Ω.

For example, using the fact that for large N, ξ_n is approximately normally distributed, that is

$$(\xi_n - p) / \sqrt{(\xi_n(1-\xi_n)/N)} \sim \text{norm}(0,1)$$

we can establish the confidence interval ($\beta = 1 - (1-\alpha)/2$)

$$p = \xi_n \pm t_\beta \sqrt{(\xi_n(1-\xi_n)/N)} \qquad (\alpha)$$

where $P(\eta \leq t_\beta) = \beta$ for a normally distributed random variable η with mean value 0 and standard deviation 1. In other words t_β is the β-fractile of the normal distribution.

Inserted into Eq.(4) this yields

$$(t_\beta \sqrt{(\xi_n(1-\xi_n) / N)}) / \xi_n < \theta$$

or with $\gamma = \theta / t_\beta$

$$N_E / N = \xi_n > 1 / (1 + \gamma^2 N)$$

to determine threshold values for N and N_E.

In a very simple version of the scheme, we could let $f(x)$ be a piecewise linear function such that $f(0) = x_1$, $f(x) = x$ and $f(1) = x_2$. Ideally $f(x)$ should be a continuous, monotonically increasing function, where $f(0) = 0$, $f(x) = x$ and $f(1) = 1$, and where the shape of $f(x)$ would be

continuously parameterized by the degree of certainty in x. A refinement of the simple version to match these criteria (omitting the degenerate cases where it does not hold that $0 < x_1 < x < x_2 < 1$ for the sake of brevity) is letting the function $f(x)$ be piece-wise linear with $f(0) = 0$, $f(t_1) = x_1$, $f(x) = x$, $f(t_2) = x_2$ and $f(1) = 1$, where say $t_1 = (1-\alpha)x$ and $t_2 = \alpha + (1-\alpha)x$.

3.2.2 Using a weighted sum

Another alternative is to use a weighted sum of the estimates:

$$p = \lambda x + \lambda' p'$$

We want λ and λ' to depend on N, the number of observations in the more specific context. We will also require that $\lambda + \lambda' = 1$. Two other desired properties are $N = 0 \Rightarrow p = p'$ and $\lim_{N \to \infty} p = x$.

A very simple strategy is to set p to p' if $N = 0$ and to x otherwise. This means that we abstract only if there are no observations at all in the specific context. We can view this as setting $\lambda(N)$ to a unit step for $N = 1$.

A less naive strategy draws inspiration from the standard deviation. Since the standard deviation behaves asymptotically as $1 / \sqrt{N}$ when N tends to infinity, we want $p - x$ to do likewise. Two different weighted sums meeting these criteria immediately spring to mind:

$$p = (\sqrt{N}\, x + p') / (\sqrt{N} + 1)$$

and

$$p = x + (p' - x) / \sqrt{(N + 1)}$$

The first one simply up-weights the specific estimate with $\sqrt{N}$, the active ingredient of the standard deviation. The second one interpolates linearly between p' and x. The distance from x is proportional to $1 / \sqrt{(N + 1)}$ (The "$+ 1$" is a technicality).

These three methods were pitted against the confidence-interval method in one of the experiments.

4 Information sources

We are now in a position to discuss how to extract morphological and syntactic information and formulate it as $P(H \mid M)$ and $P(H \mid S)$. The basic idea is to approximate these quantities with their relative frequencies. However, when this data is scarce, we will resort to the successive abstraction scheme of Section 3.

The literal ending of the word was inspected as it was suspected to contain crucial clues to the lexical category. For example, in English, any multi-syllable word ending with "-able" is almost certainly an adjective. This is even more accentuated in a language like Swedish, which has a richer inflectional and productive morphology.

To abstract, a letter was substituted with a vowel/consonant marker and abstraction was linear with earlier letters generalized before later. The last 0 – 7 letters of the words were taken into account in the experiments. The number of syllables in the rest of the word was another piece of evidence. The spectrum was zero–one–many and the single abstraction was to any number of syllables. These two generalizations competed in a greedy fashion.

The tags of the neighbouring words in the sentence were recorded. This is the conventional information source, and was believed to be very useful. However, in the experiments this information source proved much less important than the word form. Here, abstraction meant disregarding one of the neighbours at a time (the one furthest away from the word). N-gram refers to inspecting $N-1$ neighbours. Unigram through pentagram statistics were used in the experiments.

5 The experiments

The corpora used both for training and testing were portions of the Teleman corpus, a hand-tagged corpus of almost 80,000 words of miscellaneous Swedish texts (Teleman 1974). Three corpus sizes were tested in the experiments, 800, 7,500 and 65,000 words.

The tag palette used in the experiments is not Teleman's original one of around 250 different tags, but the usual set of lexical categories: Adjectives (adj), nouns, prepositions (prep), verbs, adverbs (adv), determiners (det), pronouns (pron) conjunctions (conj) and number (num), extended with proper names (name), sentence delimiters (eos), the infinitive mark "att " (inf) and characters (char), like "(", "$ " etc.

One observation of the correct hypothesis was removed when calculating its probability to simulate that this observation was not present in the training set. Important to note is that in the bulk mass of these experiments, the tags of the neighbouring words were not assigned by statistical optimization; instead the correct ones were used. This was done to allow gathering enough data to come to grips with the relative importance of the various information sources. However, for a few specific settings of the parameters, a dynamic programming technique was used to estimate the tags of the neighbouring words instead of using the pre-assigned ones.

The successive abstraction scheme employed a somewhat crude version of the confidence-interval method where the normal-distribution approximation was used for all observations except those of zero or all hits, for which the exact values from the binomial distribution were easily obtainable. The confidence level was 95 percent and the tolerance level ± 30 percent.

The task that the tagger carried out was to for each word in the corpus rank the set of tags according to the probability it assigned to them. Section 5.1 tabulates an overview of the results, while Section 5.2 examines one of the table entries in more detail, and Section 5.3 compares it with a tagging experiment were the neighbouring tags were estimated as well.

Section 5.4 compares various versions of the successive abstraction scheme.

5.1 Overview of the results

Table 1 shows an overview of the results given as token percent correct first alternatives, leading to a number of interesting conclusions.

The literal ending of the word is by far the most important information source. The neighbouring tags and the number of syllables are not at all as useful. Each extra letter seems to cost an order of magnitude in training data — the 800 word corpus peaks between 4 and 5 letters, the 7,500 word corpus between 5 and 6, and the 65,000 between 6 and 7 letters. Considering more than two neighbouring tags (i.e. using 4-gram and 5-gram statistics) improves the accuracy only marginally.

TABLE 1 : Token percent correct first alternatives.

Corpus size	Syllable information	Syntactic context	Number of final letters inspected 0	1	2	3	4	5	6	7
		1-gram	27.30	53.72	65.88	75.43	77.30	77.05	76.55	76.18
		2-gram	34.49	58.93	66.38	77.33	80.15	80.02	79.40	78.78
	any	3-gram	46.15	61.17	68.24	77.42	80.52	80.89	80.02	79.16
		4-gram	47.39	61.66	68.11	76.18	80.02	80.40	79.28	78.16
800		5-gram	47.52	62.28	67.87	76.55	79.65	80.27	79.40	78.66
words		1-gram	41.44	62.16	73.08	77.17	78.91	76.80	77.17	76.67
		2-gram	47.89	62.78	73.20	80.40	81.39	80.40	79.53	78.78
	0-1-2^+	3-gram	53.47	65.51	74.19	80.52	81.89	81.51	80.02	79.53
		4-gram	55.96	66.38	74.44	80.15	81.14	81.51	79.78	79.03
		5-gram	56.58	66.87	74.44	80.89	81.39	81.64	80.02	79.28
		1-gram	26.27	51.24	67.10	82.92	86.54	88.04	87.97	87.34
		2-gram	26.08	57.34	69.88	84.41	87.57	88.61	88.55	88.08
	any	3-gram	44.79	63.03	74.67	86.00	88.78	89.70	89.48	89.06
		4-gram	48.10	64.06	75.12	86.16	88.90	89.75	89.57	89.25
7,500		5-gram	48.39	64.33	74.79	86.01	88.90	89.72	89.72	89.40
words		1-gram	39.45	60.77	74.38	84.30	87.05	88.17	87.97	87.41
		2-gram	45.12	64.70	75.70	86.29	87.87	88.82	88.61	88.23
	0-1-2^+	3-gram	54.74	68.78	80.00	87.79	89.24	90.03	89.79	89.31
		4-gram	55.59	69.78	80.29	88.08	89.56	90.07	89.82	89.37
		5-gram	56.21	70.15	80.27	87.99	89.50	90.06	89.90	89.62
		1-gram	25.15	47.71	65.38	83.22	90.52	92.63	93.22	93.17
		2-gram	31.23	53.44	69.02	84.56	91.43	93.11	93.55	93.56
	any	3-gram	46.79	61.72	75.35	87.29	92.49	93.91	94.25	94.21
		4-gram	49.35	63.75	76.39	87.92	92.60	93.98	94.32	94.32
65,000		5-gram	51.22	64.94	76.46	88.16	92.74	94.18	94.46	??.??
words		1-gram	38.72	58.37	75.08	87.65	91.54	92.94	93.33	93.22
		2-gram	45.37	62.71	77.31	88.75	92.20	93.38	93.72	93.63
	0-1-2^+	3-gram	55.22	68.70	80.69	90.23	**93.15**	94.15	94.36	94.31
		4-gram	56.94	70.38	81.98	90.54	93.30	94.24	94.44	94.44
		5-gram	58.31	71.18	82.24	90.73	93.42	94.48	94.61	??.??

A final observation is the notorious "96 percent asymptote" reported from many statistical tagging experiments.

5.2 An expanded table entry

Table 2 shows an expanded entry from the previous table — that in boldface — where the four last letters, the number of syllable preceding those, and two neighbouring tags, were taken into account. The other entries exhibit the same general behavior.

Seeing that nouns and verbs are the most common word types, it is only reasonable that the total average should be close to the figures for these two word classes. Since the corpus is normalized, no distinction is made between capital letters and commons, and the tagger isn't doing too well on spotting names. Also, as one might expect, the tagger is having a bit of trouble telling adjectives from adverbs. A bit more surprising is that the tagger is performing so poorly on conjunctions and numbers, which are generally considered closed word classes, and should not be too difficult

to learn. The explanation to this is to be sought in the way the Teleman corpus is tagged.

TABLE 2 : 65,000 words, 3-gram syntax, 4 letters and syllable information.

Tag	1st	2nd	3rd	4-5th	6-10th	>10th	Observations
adj	85.84	10.48	2.35	1.08	0.25	0.00	4894
noun	95.39	3.43	0.83	0.33	0.03	0.00	16275
prep	98.04	1.46	0.26	0.09	0.14	0.00	7587
verb	91.71	6.18	1.30	0.61	0.20	0.00	10573
char	98.33	0.30	0.00	0.15	0.91	0.30	659
eos	99.89	0.09	0.00	0.02	0.00	0.00	4521
inf	99.47	0.46	0.00	0.00	0.08	0.00	1314
adv	88.51	7.43	2.78	1.12	0.17	0.00	4646
det	99.06	0.69	0.06	0.00	0.19	0.00	1600
pron	94.11	4.20	1.04	0.54	0.10	0.00	7184
conj	84.97	13.68	0.57	0.39	0.39	0.00	3340
num	87.07	4.36	1.15	3.06	4.13	0.23	1307
name	63.68	17.67	5.03	6.87	6.26	0.49	815
Total	93.15	4.89	1.06	0.59	0.30	0.01	64715

It is however note-worthy that the correct word class is among the two highest ranking alternatives over 98 percent of the time.

5.3 A dynamic programming version

In another version of the scheme, where dynamic programming was used to estimate the (two) neighbouring tags, rather than simply inspecting the pre-assigned ones, very similar results were recorded. This fact lends further strength to that claim that morphological information is of much greater importance than the local syntactic context.

A few settings of the various parameters were tested using this scheme, all yielding results conforming to those of Table 3, where the 65,000 word corpus was used, and the four last letters and the number of preceding syllables were employed as morphological information sources. The figure given is again token percent correct first alternatives.

5.4 Varying the successive abstraction scheme

Four different schemes for combining the accurate estimate p' from the general context with the potentially inaccurate estimate x from the specific context were tried out using 3-gram local syntactic information (i.e. two neighbouring tags), and inspecting the four final letters ot the word and the number of preceeding syllables.

TABLE 3 : Comparison between knowing and guessing neighbour tags.

Tag	Known tags	Guessed tags	No tags	Observations
adj	85.84	86.31	83.27	4894
noun	95.39	95.43	92.11	16275
prep	98.04	97.72	98.00	7587
verb	91.71	90.93	89.97	10573
char	98.33	97.57	98.63	659
eos	99.89	97.01	99.87	4521
inf	99.47	98.78	99.85	1314
adv	88.51	87.88	87.77	4646
det	99.06	98.94	98.88	1600
pron	94.11	91.65	95.16	7184
conj	84.97	85.99	79.61	3340
num	87.07	87.83	84.85	1307
name	63.68	65.15	59.63	815
Total	93.15	92.59	91.54	64715

The four strategies were:[1]

1. The confidence interval method as described above.

2. $\boldsymbol{p} = \boldsymbol{p}'$ if $N = 0$,
   $\boldsymbol{p} = \boldsymbol{x}$ if $N > 0$.
   We abstract only if there is no data available at all.

3. $\boldsymbol{p} = (\sqrt{N}\,\boldsymbol{x} + \boldsymbol{p}') / (\sqrt{N} + 1)$.
   The weight of the specific result is simply $\sqrt{N}$ and the sum is normalized.

4. $\boldsymbol{p} = \boldsymbol{x} + (\boldsymbol{p}' - \boldsymbol{x}) / \sqrt{(N + 1)}$.
   The result is on the line between the specific and the general estimate. The distance from the specific eatimate is proportional to $1 / \sqrt{(N + 1)}$.

The results shown in Table 4 reveal that the last two strategies, the weighted-sum methods, are quite superior to the first two, the first one of them being the slightly better. Strategy 1, the confidence-interval method, is only somewhat better than not abstracting at all until forced to, as is done in strategy 2, when both syntactic and morphological information is taken into account.

The explanation for this could be the following: Even though data might be scarce, what is there is there, and those particular observations are more likely to be there as a result of having a higher probability, than by pure chance.

[1] Again N is the total number of observations in the specific context.

With only 800 words, the data can safely be assumed to be scare and the successive abstraction scheme improves the parameter estimates considerably. This is especially true when syntactic and morphological information is combined, and something less coarse grained than a mere ranking of the alternatives is required. Already with 7,500 words, though, the improvements are small and for 65,000 words, where one would expect sufficient data to be available for most estimates, the improvements are marginal. However, at least strategy 3 does not seem to degrade performance.

The best result observed, 95.38 percent, was for the setting of 6 letters, syllable information, 4-gram syntax (three neighbouring tags) and strategy 3 on the 65,000 word corpus.

TABLE 4 : Comparison between different successive abstraction schemes.

Corpus	Strategy	1	2	3	4
800 words	Syntax and morphology	81.89	78.29	86.35	85.86
	Morphology only	78.91	83.62	84.49	84.12
	Syntax only	46.15	48.88	46.03	46.53
7,500 words	Syntax and morphology	88.78	88.74	91.14	90.94
	Morphology only	87.05	90.07	89.83	89.48
	Syntax only	44.79	45.90	45.57	45.82
65,000 words	Syntax and morphology	93.15	93.83	94.06	93.78
	Morphology only	91.54	92.90	92.76	92.46
	Syntax only	46.79	46.80	46.87	46.89

6 Summary and conclusions

A number of interesting results emerged from these experiments. Even though it is not very surprising that the literal appearance of a word is a much more important information source than its local syntactic context for assigning the correct lexical category, it *is* surprising that it is so much more important. The *global* syntactic context, on the other hand, has proved very useful as reported in (Voutilainen *et al* 1992).

The design of the tagger relies heavily on the successive-abstraction scheme. The results are a success for the scheme even though it is a bit disappointing that the simpler weighted-sum method out-performed the more elaborate confidence-interval method. The moral might be phrased "If one wants a point estimate, one shouldn't stare too intensely at confidence intervals".

One tends to consider the Teleman corpus is a bit oddly tagged seeing that the tagger is having difficulties assigning the correct tag to closed class words such as conjunctions, numbers and pronouns.

Finally, the peak performance value of the tagger, 95.38 token percent correct first alternatives, is quite respectable in itself. However, two other approaches to the same task indicate that this result can be improved on. Cutting (1994) attempts the same task by using a lexicon and an untagged corpus to train from, making predictions using only bigram syntactic information in addition to lexical probabilities, and reports 95 percent success rate. This is probably a somewhat more difficult task. Eineborg and Gambäck (1994) report a success rate of 96.3 percent using a neural net with 4-gram statistics and six letter endings. They employ an intermediate abstraction level based on grouping the letters into phonological classes such as fricatives, explosives etc. This could readily be incorporated into the scheme described in this paper and could potentially improve its performance.

Acknowledgements

This work was made possible by the basic research programme at the Swedish Institute of Computer Science (SICS). I would very much like to thank my friend and colleague Jussi Karlgren for inspiring discussions and for pointing me to related work, Gunnel Källgren at Stockholm University for support and encouragement, Krister Lindén at Helsinki University for introducing me to the excellent work by Finnish researchers in this field and Dave Carter and Manny Rayner at SRI International, Cambridge, for comments and valuable suggestions to improvements on draft versions of this paper. I enjoyed very much the friendly competition from Douglass Cutting, Martin Eineborg and Björn Gambäck and the opportunity to discuss and compare our different approaches. I finally thank my other colleagues at SICS for contributing to a very conductive atmosphere to work in.

References

Brill, Eric. 1992. *A Simple Rule-Based Part of Speech Tagger.* pp. 152–155 of PROCS. 3RD ANLP , Trento, Italy.

Church, Kenneth W. 1988. *A Stochastic Parts of Speech and Noun Phrase Parser for Unrestricted Text.* pp. 136–143 of PROCS. 2ND ANLP.

Cutting, Douglass. 1993. *A Practical Part-of-Speech Tagger.* In Eklund (ed.), *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3–5 June 1993*, Stockholm, Sweden.

DeRose, Steven J. 1988. *Grammatical Category Disambiguation by Statistical Optimization.* pp. 31–39 of COMPUTATIONAL LINGUISTICS, Volume 14, Number 1.

Eklund, Robert, (ed.). 1994. *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3–5 June 1993*, Stockholm, Sweden.

Eineborg, Martin and Björn Gambäck. 1994. *Tagging Experiments Using Neural Networks.* In Eklund (ed.), *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3–5 June 1993*, Stockholm, Sweden.

Källgren, Gunnel. 1991. *Parsing without lexicon: the MorP system.* PROCS. of 5TH EACL, Berlin, Germany.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, California.

Teleman, Ulf. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska* (in Swedish), Studentlitteratur, Lund, Sweden.

Voutilainen, Atro, Juha Heikkilä and Arto Anttila. 1992. *Constraint Grammar of English.* Publication Number 21, Deptartment of General Linguistics, University of Helsinki, Finland.