Using Word Class for Part-of-speech Disambiguation

Evelyne Tzoukermann and Dragomir R. Radev*
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974–0636
evelyne,s_radev@research.att.com

*Department of Computer Science
Columbia University
New York, NY 10027
radev@cs.columbia.edu

Abstract

This paper presents a methodology for improving part-of-speech disambiguation using word classes. We build on earlier work for tagging French where we showed that statistical estimates can be computed without lexical probabilities. We investigate new directions for coming up with different kinds of probabilities based on paradigms of tags for given words. We base estimates not on the words, but on the **set of tags** associated with a word. We compute frequencies of unigrams, bigrams, and trigrams of word classes in order to further refine the disambiguation. This new approach gives a more efficient representation of the data in order to disambiguate word part-of-speech. We show empirical results to support our claim. We demonstrate that, besides providing good estimates for disambiguation, word classes solve some of the problems caused by sparse training data. We describe a part-of-speech tagger built on these principles and we suggest a methodology for developing an adequate training corpus.

1 Introduction

In the part-of-speech literature, whether taggers are based on a rule-based approach (Klein and Simmons, 1963), (Brill, 1992), (Voutilainen, 1993), or on a statistical one (Bahl and Mercer, 1976), (Leech et al., 1983), (Merialdo, 1994), (DeRose, 1988), (Church, 1989), (Cutting et al., 1992), there is a debate as to whether more attention should be paid to lexical probabilities rather than contextual ones. (Church, 1992) claims that part-of-speech taggers depend almost exclusively on lexical probabilities, whereas other researchers, such as Voutilainen (Karlsson et al., 1995) argue that word ambiguities vary widely in function of the specific text and genre. Indeed, part of Church's argument is relevant if a system is based on a large corpus such as the Brown corpus (Francis and Kučera, 1982) which represents one million surface forms of morpho-syntactically disambiguated words from a range of balanced texts. Consider, for example, a word like "cover" as discussed by Voutilainen (Karlsson et al., 1995): in the Brown and the LOB Corpus (Johansson, 1980), the word "cover" is a noun 40% of the occurrences and a verb 60% of the other, but in the context of a car maintenance manual, it is a noun 100% of the time. Since, for statistical taggers, 90% of texts can be disambiguated solely applying lexical probabilities, it is, in fact, tempting to think that with more data and more accurate lexical estimates, more text could be better disambiguated. If this hypothesis is true for English, we show that it does not hold for languages for which publicly available tagged corpora do not exist. We also argue against Church's position, supporting the claim that more attention needs to be paid to contextual information for part-of-speech disambiguation (Tzoukermann et al., 1995).

The problem tackled here is to develop an "efficient" training corpus. Unless large effort, money, and time are devoted to this project, only small corpora can be disambiguated manually. Consequently, the problem of extracting lexical probabilities from a small training corpus is twofold: first, the statistical model may not necessarily represent the use of a particular word in a particular context. In a morphologically inflected language, this argument is particularly serious since a word can be tagged with a large number of parts of speech, i.e. the ambiguity potential is high. Second, word ambiguity may vary widely depending on the particular genre of the text, and this could differ from the training corpus. When there is no equivalent for the Brown corpus in French, how should one build an adequate training corpus which reflects properly lexical probabilities? How can the numerous morphological variants that render this task even harder be handled?

The next section gives examples from French and describes how morphology affects part-of-speech disambiguation and what types of ambiguities are found in the language. Section 3 examines different techniques used to obtain lexical probabilities. Given the problems created by estimating probabilities on a corpus of restricted size, we present in Section 4 a solution for coping with these difficulties. We suggest a new paradigm called **genotype**, derived from the concept of ambiguity class (Kupiec, 1992), which gives a more efficient representation of the data in order to achieve more accuracy in part-of-speech disambiguation. Section 5 shows how our approach differs from the approach taken by Cutting and Kupiec. The frequencies of unigram, bigram, and trigram genotypes are computed in order to further refine the disambiguation and results are provided to support our claims. The final section offers a methodology for developing an adequate training corpus.

2 French words and morphological variants

To illustrate our position, we consider the case of French, a typical Romance language. French has a rich morphological system for verbs – which can have as many as 48 inflected forms – and a less rich inflectional system for nouns and adjectives, the latter varying in gender and number having up to four different forms. For example, the word "marine" shown in Table 1, can have as many as eight morphological analyses.

word	**base form**	**morphological analysis**	**tags**
"marine"	<marin>	adjective, feminine singular	jfs
"marine"	<marine>	noun, feminine singular	nfs
"marine"	<marine>	noun, masculine singular	nms
"marine"	<mariner>	verb, 1st person, singular, present, indicative	v1spi
"marine"	<mariner>	verb, 1st person, singular, present, subjunctive	v1sps
"marine"	<mariner>	verb, 2nd person, singular, present, imperative	v2spm
"marine"	<mariner>	verb, 3rd person, singular, present, indicative	v3spi
"marine"	<mariner>	verb, 3rd person, singular, present, subjunctive	v3sps

Table 1: Morphological analyses of the word "marine".

The same word "marine", inflected in all forms of the three syntactic categories (adjective, noun, and verb) would have 56 morphologically distinct forms, i.e. 4 for the adjective, 2 for each of the nouns, and 48 for the verb. At the same time, if we collapse the homographs, these 56 morphologically distinct forms get reduced to 37 homographically distinct forms and the ambiguity lies in the 19 forms which overlap across internal verb categories, but also across nouns and adjectives. Table 1 shows 5 verb ambiguities, 2 noun ambiguities, a total of 8 homographs including the adjective form.

Part-of-speech Ambiguity of French words. Once morphological analysis is completed, ambiguity of words is computed in order to locate the difficulties. Figure 1 shows two corpora of different sizes and the number of words each tag contains. The figure clearly exhibits that even though Corpus 2 is twice as large as Corpus 1, the distribution of words per tags is very similar, i.e. more than 50% of the words have only one tag and are thus unambiguous, 25% of the words have two tags, 11% of the words have three tags, and about 5% of the words have from four to eight tags.

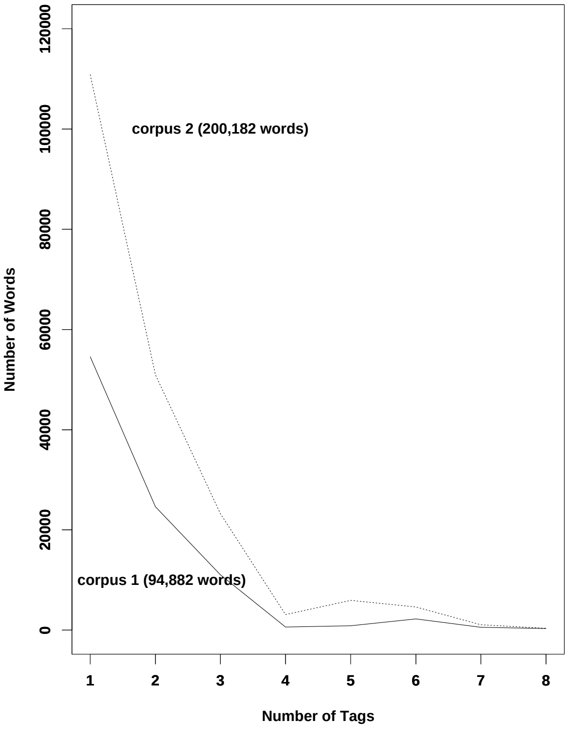

Figure 1: Number of words per ambiguity level in two different corpora

3 Problems with lexical probabilities

There are several ways lexical probabilities could be estimated for a given language, each of them presenting problems:

1. **From raw text**: a human tagger could manually disambiguate texts. There are some problems though due to the fact that there are always words that are overseen (therefore improperly tagged) or there is disagreement between humans (on at least 5% of the words),

and cross-checking by another human is required. In our system, we manually tagged about 76,000 words[1] in this way.

2. **Bootstrapping from already tagged text**: this technique generally consists of using a small tagged corpus to train a system and having the system tag another subset of the corpus that gets disambiguated later. (Derouault and Merialdo, 1986) have used these techniques but the necessary human effort is still considerable.

3. **From the baseform of the word**: one could estimate the frequency of the analyzed stem in the process of morphological analysis.

4. **From the inflectional morpheme**: similarly, one could estimate the probability of the inflectional morpheme given its stem. This approach is often used for smoothing probabilities, but, considering the high ambiguity of some French suffixes, such as "e", "es", etc, it is doubtful that basing the estimates on the suffixes alone would give good results.

5. **From unseen pairs of [words,tags]**: for a given word, such as "marine" that can have 8 possible tags, if only the instances [marine, adj-fem-sing], [marine, noun-fem-sing] are found in the training corpus, one could assume that the remaining unseen instances have a much lower probability. This could create problems in making incorrect assumptions on words.

Out of all the possibilities outlined above, none seems feasible and robust enough. Therefore, we decided to pay more attention to a different paradigm which captures more information about the word at a morphological and syntactic level.

4 The genotype solution

In an attempt to capture the multiple word ambiguities on the one hand and the recurrence of these observations on the other, we came up with a new concept, called **genotype**. In biology, the genotype refers to the content of genes or the pattern of genes in the cell. As used in our context, the genotype is the set of part of speech tags associated with a word. Each word has a genotype (or series of tags based on morphological features) assigned during morphological analysis, and words, according to their patterns, share the same genotype. The genotype depends on the tagset, but not on any particular tagging method. For example, the word "marine" with the eight morphological analyses listed in Table 1, has the genotype [JFS NFS NMS V1SPI V1SPS V2SPM V3SPI V3SPS][2], each tag corresponding to an analysis, i.e. the list of potential tags for "marine" as shown in Table 1. For each genotype, we compute the frequency with which each of the tags occurs and we select this decision. This paradigm has the advantage of capturing the morphological variation of words combined with the frequency with which they occur. A **genotype decision** is the most frequent tag associated with a genotype in the training corpus. As explained in Section 4.2, out of a training corpus of 76,000 tokens, we extracted a total of 429 unigram genotypes, 6650 bigram genotypes, and 23,802 trigram genotypes with their respective decisions.

[1] We wish to thank Anne Abeillé and Thierry Poibeau for helping the manual tagging.

[2] JFS = adjective, feminine, singular; NFS = noun, feminine, singular; NMS = noun, masculine, singular; V1SPI = verb, 1st person, singular, present, indicative; V1SPS = verb, 1st person, singular, present, subjunctive; V2SPM = verb, 2nd person, singular, present, imperative; V3SPI = verb, 3rd person, singular, present, indicative; V3SPS = verb, 3rd person, singular, present, subjunctive.

4.1 Power of genotypes

The genotype concept allows generalizations to be made across words according to tag patterns, thereby gathering estimates not on words but on tag occurrences. We discovered that in a training corpus of 76,000 tokens, lexical frequencies are not as reliable as genotype frequencies. In order to illustrate this, Table 2 and Table 3 show convincing results using this approach. Table 2 presents the set of words corresponding to the genotype [NFP V2S], and their resolution with respect to lexical frequencies and genotype frequencies. The table shows 12 words from the test corpus which, from a morphological point of view, can be either verb-2nd-person-singular (V2S) or noun-feminine-plural (NFP); the first column contains always the same tag NFP, because of the genotype decision; we learned from the training corpus that at each time a word could be tagged NFP or V2S, it is 100% of the times NFP, 0% V2S, therefore the noun form is always picked over the verb form. Out of the 12 words listed in the Table 2, 4 words (marked *unseen* in the table) could not be estimated using lexical frequencies alone since they do not appear in the training corpus. However, since all of them belong to the same genotype, the 4 unseen occurrences are properly tagged.

	genotype decision	**lexical decision**	**correct decision**
oeuvres	nfp	*unseen*	nfp
dépenses	nfp	nfp	nfp
dépenses	nfp	nfp	nfp
toiles	nfp	*unseen*	nfp
affaires	nfp	nfp	nfp
avances	nfp	*unseen*	nfp
finances	nfp	nfp	nfp
feuilles	nfp	nfp	nfp
forces	nfp	nfp	nfp
oeuvres	nfp	*unseen*	nfp
tâches	nfp	nfp	nfp
réformes	nfp	nfp	nfp

Table 2: [NFP V2S] genotype frequencies vs lexical frequencies

In Table 3, we demonstrate that the genotype decision for the [NMS V1S V2S V3S] genotype always favors the noun-masculine-singular form (NMS) over the verb forms (V1S for verb-1st-person-singular, V2S for verb-2nd-person-singular, V3S for verb-3rd-person-singular). Out of the 12 words listed in Table 3, 5 do not occur in the training corpus and 4 of them can be properly tagged using the genotype estimates. The word "suicide", however, which should be tagged as a verb, was improperly tagged as a noun. Note that we are only considering unigrams of genotypes, which tend to overgeneralize. However, as shown in Section 4.3, the additional estimates of bigrams and trigrams will use the context to select a more appropriate tag.

4.2 Distribution of genotypes

Among all parts of speech, there is a clear division between closed-class parts of speech, which include prepositions and conjunctions, and open-class ones, which includes verbs, nouns, and adjectives. Similarly, we suggest that genotypes be classified in categories:

- **Closed-class genotypes** contain at least one closed-class part-of-speech, e.g., "des", which belongs to the [P R] (preposition, article) genotype.

	genotype decision	**lexical decision**	**correct decision**
suicide	nms	*unseen*	v3s
chiffre	nms	nms	nms
escompte	nms	*unseen*	nms
escompte	nms	*unseen*	nms
cercle	nms	*unseen*	nms
doute	nms	nms	nms
nombre	nms	nms	nms
avantage	nms	nms	nms
pilote	nms	nms	nms
peigne	nms	*unseen*	nms
doute	nms	nms	nms
groupe	nms	nms	nms

Table 3: [NMS V1S V2S V3S] genotype frequencies vs lexical frequencies

- **Semi closed-class genotypes** contain only open-class parts-of-speech, but behave very similarly to the closed-class genotype, with respect to the small number of words – often homograph – in that genotype. For instance, the word "fils" (*son* [singular and plural], *threads*) with the low frequent genotype [NM NMP] or the word "avions" (*planes, (we) had*) which belong to the genotype [NFP V1P].
- **Open-class genotypes** contain all other genotypes, such as [NFS V1S V2S V3S]. This class, unlike the other two, is productive.

There are several facts which demonstrate the power of genotypes for disambiguation. First, the number of genotypes on which the estimates are made is much smaller than the number of words on which to compute estimates. Our results show that in the training corpus of 76,000 tokens, there are 10,696 words, and 429 genotypes. Estimating probabilities on 429 genotypes rather than 10,696 words is an enormous gain. Since the distributions in both cases have a very long tail, there are many more words than genotypes for which we cannot obtain reliable statistics. As an example, we extracted the most frequent open-class genotypes from the training corpus (each of them occurring more than 100 times) shown in Table 4. It is striking to notice that these 22 genotypes represent almost 10% of the corpus. The table shows the genotype in the first column, the number of occurrences in the second one, the part-of-speech distribution in the third one, the best genotype decision and the percent of this selection in the last column. We can see that words belonging to the same genotype are likely to be tagged with the same tag; for example, the genotype [NFS V1S V2S V3S] is tagged as NFS. That allows us to make predictions for words missing from the training corpus.

4.3 Contextual probabilities via bigram and trigram genotypes

Using genotypes at the unigram level tends to result in overgeneralization, due to the fact that the genotype sets are too coarse. In order to increase the accuracy of part-of-speech disambiguation, we need to give priority to trigrams over bigrams, and to bigrams over unigrams.

In a way similar to decision trees, Table 5 shows how the use of context allows for better disambiguation of genotype. We have considered a typical ambiguous genotype [JMP NMP] which occurs 607 times in the training corpus, almost evenly distributed between the two alternative

genotype	# of occ.	distribution	decision
nfs v1s v2s v3s	899	nfs(797) v1s(0) v2s(0) v3s(100)	nfs(88.7%)
jms nms	734	jms(498) nms(230)	jms(67.8%)
jmp nmp	607	nmp(291) jmp(316)	jmp(52.6%)
nms v3s	612	nms(28) v3s(584)	v3s(95.4%)
nfp v2s	441	nfp(437) v2s(1)	nfp(99.1%)
jfs nfs	401	jfs(333) nfs(67)	jfs(83.0%)
nms v1s v2s v3s	405	nms(351) v1s(0) v2s(0) v3s(51)	nms(86.7%)
nms qsms	325	nms(52) qsms(271)	qsms(83.4%)
jfp nfp	292	jfp(192) nfp(99)	jfp(65.8%)
v1s v2s v3s	263	v1s(3) v2s(0) v3s(259)	v3s(98.5%)
nmp v2s	259	nmp(254) v2s(1)	nmp(98.1%)
nms v	249	nms(21) v(228)	v(91.6%)
jms qsms	222	jms(24) qsms(197)	qsms(88.7%)
jms nms qsms	213	jms(19) nms(33) qsms(161)	qsms(75.6%)
jfs nfs qsfs	169	jfs(8) nfs(110) qsfs(51)	nfs(65.1%)
nfs nms	131	nfs(67) nms(64)	nfs(51.1%)
nfs nms v1s v2s v3s	115	nfs(39) nms(49) v1s(0 v2s(0) v3s(27)	nms(42.6%)
jfp nfp qsfp	126	jfp(1)2 nfp(55) qsfp(58)	qsfp(46.0%)
jms nms qsms v3s	114	jms(2) nms(18) qsms(52) v3s(42)	qsms(45.6%)
jfs nfs v1s v2s v3s	110	jfs(39) nfs(27) v1s(1) v2s(0) v3s(42)	jfs(38.2%)
jmp qsmp	112	jmp(8) qsmp(103)	qsmp(91.2%)
jmp nmp qsmp	100	jmp(8) nmp(47) qsmp(45)	nmp(47.0%)

Table 4: The most frequent open-class genotypes

tags, JMP and NMP. As a result, if only unigram training data is used, the best candidate for that genotype would be JMP, occurring 316 out of 607 times. However, choosing JMP only gives us 52.06% accuracy. Table 5 clearly demonstrates that the contextual information around the genotype will bring this percentage up significantly. As an example, let us consider the 5th line of Table 5, where the number 17 is marked with a square. In this case, we know that the [JMP NMP] genotype has a right context consisting of the genotype [p r] (4th column, 5th line). In this case, it is no longer true that JMP is the best candidate. Instead, NMP occurs 71 out of 91 times and becomes the best candidate. Overall, for all possible left and right contexts of [JMP NMP], the guess based on both the genotype and the single left or right contexts will be correct 433 times out of 536 (or 80.78%). In a similar fashion, the three possible trigram layouts (Left, Middle, and Right) are shown in lines 18-27. They show that the performance based on trigrams is 95.90%. This particular example provides strong evidence of the usefulness of contextual disambiguation with genotypes. The fact that this genotype, very ambiguous as a unigram (52.06%), can be disambiguated as a noun or adjective according to context at the trigram stage with 95.90% accuracy demonstrates the strength of our approach.

4.4 Smoothing probabilities with genotypes

In the context of a small training corpus, the problem of sparse data is more serious than with a larger tagged corpus. Genotypes play an important role for smoothing probabilities. By paying attention to tags only and thus ignoring the words themselves, this approach handles new words that have not been seen in the training corpus. Table 6 shows how the training corpus provides coverage for n-gram genotypes that appear in the test corpus. It is interesting to notice that only

n-gram	pos.	total	genotype	decision	distr.	cor.	total	cor.	total	accuracy
Unigram		607	**[jmp nmp]**	**jmp**	**316**	316	607	316	607	52.06%
				nmp	291					
Bigram	Left	230	**[jmp nmp]**[x]	**jmp**, x	**71**	71	102	433	536	80.78%
				nmp, x	31					
			[jmp nmp][p r]	**jmp**, p	**17**	71	91			
				jmp, r	3					
				nmp, p	71					
			[jmp nmp][nmp]	**jmp**, nmp	**23**	23	24			
				nmp, nmp	1					
			[jmp nmp][a]	**jmp**, a	**13**	13	13			
	Right	306	[p r]**[jmp nmp]**	p, **jmp**	27	112	141			
				p, **nmp**	**104**					
				r, **jmp**	2					
				r, **nmp**	**8**					
			[b r]**[jmp nmp]**	r, **jmp**	22	72	94			
				r, **nmp**	**72**					
			[nmp]**[jmp nmp]**	nmp, **jmp**	**71**	71	71			
Trigram	Left	32	**[jmp nmp]**[p r][nms]	**nmp**, p, nms	**21**	21	21	117	122	95.90%
			[jmp nmp][jmp nmp][x]	**jmp**, jmp, x	3	8	11			
				nmp, jmp, x	**8**					
	Middle	44	[p r]**[jmp nmp]**[p r]	p, **nmp**, p	**23**	23	23			
			[b r]**[jmp nmp]**[p r]	r, **nmp**, p	**19**	19	21			
				r, **jmp**, p	2					
	Right	46	[p r][nmp]**[jmp nmp]**	p, nmp, **jmp**	**27**	29	29			
				r, nmp, **jmp**	**2**					
			[n z][p r]**[jmp nmp]**	z, p, **nmp**	**16**	17	17			
				z, r, **nmp**	**1**					

Table 5: Influence of context for n-gram genotype disambiguation.

12 out of 1564 unigram genotypes (0.8%) are not covered. The training corpus covers 71.4% of the bigram genotypes that appear in the test corpus and 22.2% of the trigrams.

Coverage of Genotypes			
	test corpus	**training corpus**	
	# of genotypes	**# of genotypes**	%
1-grams	1564	1552	(99.2 %)
2-grams	1563	1116	(71.4 %)
3-grams	1562	346	(22.2 %)

Table 6: Coverage in the training corpus of n-gram genotypes that appear in the test corpus.

5 Comparison with other approaches

In some sense, this approach is similar to the notion of "ambiguity classes" explained in (Kupiec, 1992) and (Cutting et al., 1992) where words that belong to the same part-of-speech figure together. In this approach, they use the notion of word equivalence or ambiguity classes to describe words belonging to the same part-of-speech categories. In our work, the entire algorithm bases estimations on genotype only, filtering down the ambiguities and resolving them with statistics. Moreover, the estimation is achieved on a sequence of n-gram genotypes. Also, the refinement that is contained in our system reflects the real morphological ambiguities, due to the rich nature of the morphological output and the choice of tags. There are three main differences between their work and ours. First, in their work, the most common words are estimated individually and the less common ones are put together in their respective ambiguity classes; in our work, every word is equally treated by its respective genotype. Second, in their work, ambiguity classes can be marked with a preferred tag in order to help disambiguation whereas in our work, there is no special annotation since words get disambiguated through the sequential application of the modules. Third, and perhaps the most important, in our system, the linguistic and statistical estimations are entirely done on the genotypes only, regardless of the words. Words are not estimated individually given their class categories; rather, genotypes are estimated separately from the words or in the context of other genotypes (bi- and tri-gram probabilities). (Brill, 1995) presents a rule-based part-of-speech tagger for unsupervised training corpus. Some of the rules of his system and the fact that he uses a minimal training corpus suggests some similarities with our system, but the main aim of the work is to investigate methods to combine supervised and unsupervised training in order to come up with a highly performing tagger. (Chanod and Tapanainen, 1995) compare two tagging frameworks for tagging French, one that is statistical, built upon the Xerox tagger (Cutting et al., 1992), and another based on linguistic constraints only. The contraints can be 100% accurate or describe the tendency of a particular tagging choice. The contraint-based tagger is proven to have better performance than the statistical one, since rule writing is more handlable or more controllable than adjusting the parameters of the statistical tagger. It is difficult to compare any kind of performance since their tagset is very small, i.e. 37 tags, including a number of word-specific tags (which reduces further the number of "real" tags), and does not account for several morphological features, such as gender, number for pronouns, etc. Moreover, categories that can be very ambiguous, such as coordinating conjunctions, subordinating conjunctions, relative and interrogative pronouns which tend to be collapsed; consequently, the disambiguation is simplified and results cannot be compared.

6 Implementation and performance of the part-of-speech tagger

We have developed a part-of-speech tagger using only a finite-state machine framework. The input string is represented as a finite-state generator, and the tagging is obtained through composition with a pipeline of finite-state transducers (FST's). Besides the modules for pre-processing and tokenization, the tagger includes a morphological FST and a statistical FST, which incorporates linguistic and statistical knowledge. We have used a toolkit developed at AT&T Bell Laboratories (Pereira et al., 1994) which manipulates weighted and unweighted finite-state machines (acceptors or transducers). Using these tools, we have created a set of programs which generate finite-state transducers from descriptions of linguistic rules (in the form of negative constraints) and for encoding distribution information obtained through statistical learning. Statistical decisions on genotypes are represented by weights – the lower cost, the higher the chance of a particular tag to be picked. With this representation, we are able to prefer one n-gram decision over another based on the cost.

The morphological FST is generated automatically from a large dictionary of French of about 90,000 entries and on-line corpora, such as Le Monde Newspapers (of the European Community Initiative, 1989 and 1990). It takes the text as input and produces an FST that encodes each possible tagging of the input text as one distinct path from the start state to the final state. The statistical FST is created from 1-gram, 2-gram, and 3-gram genotype data obtained empirically from the training corpus. It encodes all 1, 2, 3-grams of genotypes extracted from the training corpus with a cost determined as a function of the frequency of the genotype decision in the training corpus. Table 7 shows how costs are computed for a specific bigram and how these costs are used to make a tagging decision. The bigram in the example, [p r] [jmp nmp], occurs 306 times in the training corpus. All possible taggings, i.e. [p] [jmp], [p] [nmp], [r] [jmp], and [r] [nmp] appear in the training corpus. The sub-FST that corresponds to this bigram of genotypes will have [p r] [jmp nmp] on its input and all 4 possible taggings on its output. Each tagging sequence has a different costs. Let f be the total count of the genotype bigram. Let f_t be the number of cases that the tagging is t, for all possible taggings t (in this example there are 4 possible taggings). The cost of the transition for tagging t is the negative logarithm of f_t divided by f: $-\log(f_t/f)$. The selected transition is the one with the lowest cost; the example in Table 7 illustrates the computation of costs with [p] [nmp], the selected tagging in bold.

genotype bigram	tagging	frequency	cost
[p r] [jmp nmp]	p, jmp	27/306	2.43
	p, nmp	**104/306**	**1.08**
	r, jmp	2/306	5.03
	r, nmp	8/306	3.64

Table 7: An example of cost computation for the bigram FST [p r] [jmp nmp].

In a similar way, the statistical FST contains paths for unigrams and trigrams. In order to prefer trigrams over bigrams, and bigrams over unigrams, we have added a *biased cost* to some transitions. The empirically determined values of the *biased cost* are as follows:

trigram biased cost < *bigram biased cost* < *unigram biased cost.*

If a certain bigram or trigram does not appear in the training corpus, the FST will still have a corresponding path, but at a higher cost. Since negative constraints (such as "article" followed by "verb") reflect n-grams that are impossible linguistically and therefore have an expected frequency of appearance equal to 0, we assign them a very high cost (note that in order to keep the graph connected, we cannot assign a cost of ∞). To make the use of *biased cost* clear, Table 8 shows the unigrams [p r] and [jmp nmp] that compose the bigram described in Table 7 and the corresponding transition costs.

genotype unigram	tagging	frequency	cost	biased cost
[p r]	**p**	**6645/6883**	**0.04**	**1.04**
	r	238/6883	3.36	4.36
[jmp nmp]	**jmp**	**316/607**	**0.65**	**1.65**
	nmp	291/607	0.73	1.73

Table 8: An example of biased cost for the unigram sub-FST's [p r] and [jmp nmp].

Figure 2 presents the FST that corresponds to Table 7 and Table 8. The top part shows how the genotype bigram [p r] [jmp nmp] can be tagged as a sequence of two unigrams; the bottom part uses one bigram to tag it. The notation on all arcs in the FST is the following:

input string : output string / cost

e.g.,

[p n] : p / 1.04

The input is a genotype n-gram, the output represents a possible tag n-gram with the corresponding cost. The FST shown in Figure 2 is part of a much larger FST containing 2.8 million arcs.

The cheapest path for tagging the sequence of two genotypes [p r] [jmp nmp] can go either

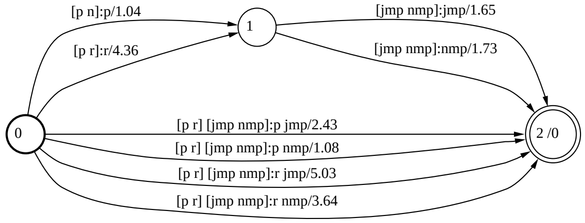
[p n]:p/1.04
[p r]:r/4.36
1
[jmp nmp]:jmp/1.65
[jmp nmp]:nmp/1.73
0
[p r] [jmp nmp]:p jmp/2.43
[p r] [jmp nmp]:p nmp/1.08
[p r] [jmp nmp]:r jmp/5.03
[p r] [jmp nmp]:r nmp/3.64
2 /0

Figure 2: Example of an FST that tags the genotype bigram [p r] [jmp nmp]

through one bigram transition shown in bold face in Table 7, or through two adjacent unigram transitions shown in bold face in Table 8. The corresponding paths through the FST are shown in Figure 2. In the first case (bigrams), the tagging of [p], [nmp] is at a cost of 1.08, whereas in the other case (unigrams), the cheapest path or the lowest cost includes the two transitions [p] and [jmp] for a total cost of $1.04 + 1.65 = 2.69$. In this case, not only do bigrams have precedence over unigrams, but the choice of the tagging sequence [p], [nmp] is also better than the sequence [p] [jmp], as it takes into account the context information. Similarly, if a trigram contained a bigram as a sub-FST, typically the cost of going through the trigram would be smaller than the cost of going through a bigram and a unigram. In the case where two consecutive genotype unigrams do not compose a bigram seen in the training corpus, there is no context information that can be applied and only the information of the tagging of the individual unigrams is used.

The tagger is based on a tagset of 72 parts of speech. As said earlier, the training corpus was manually tagged and contained 76,000 words. The test corpus, also manually tagged, contained 1,500 words. Taking into account the large number of parts of speech, the tagger disambiguates correctly about 95% of unrestricted text. We are in the process of improving the tagger performance in refining rules and biased costs.

7 Steps for building an optimal training corpus

This section explains the motivations of our claims for developing taggers for a language. The following steps are based on our experience and, we believe, will extend to a wide range of language types.

1. **Study morpho-syntactic ambiguity and word frequencies:** Part-of-speech ambiguities must be observed as a function of the word frequencies as shown in Section 2.

2. **Analyze morphology and morphological features** in order to evaluate the ambiguity of the language. As shown in Section 2, some suffixes may disambiguate a certain number of words, whereas others may be truly ambiguous and overlap over several categories of words.

3. **Determine concise tagset** based on trade-off between tagset size and computational complexity. This requires system tuning and is often dependent on the application. The more tags, the harder the estimation of probabilities, and the sparser the data. Having a concise set of tags is therefore a priority.

4. **Obtain maximum genotype coverage:** genotypes must first be separated into closed, semi-closed, and open class. Then, the first two classes must be exhaustively covered since their number is relatively small. Last, open-class genotypes should be examined by order of frequency; since their number is finite, they can also be exhaustively covered.

5. **Capture contextual probabilities:** genotypes must be considered in context. As described in Section 4.3, bigram and trigram genotypes give accurate estimates of the morpho-syntactic variations of the language.

We believe that concentrating efforts on these issues will allow part-of-speech tagger developers to optimize time and effort in order to develop adequate basic training material.

8 Conclusion

We explored the morpho-syntactic ambiguities of a language, basing our experiments on French. Several ways to estimate lexical probabilities were discussed and a new paradigm, the genotype, was presented. This paradigm has the advantage to capture the morphological variation of words along with the frequency at which they occur. A methodology is presented in order to optimize the construction of a restricted training corpus for developing taggers. In order to disambiguate word part-of-speech with a small training corpus, genotypes turn out to be much easier to model than the words themselves. They offer a successful solution to the small training corpus problem as well as to the problem of data sparsness. Compared to lexical probabilities, they give much more reliable accounts, since only 429 genotypes need to be estimated instead of 10,696 words for lexical probabilities. Results are even more convincing when genotypes are used in context and bigrams and trigrams are applied to disambiguate. Additionally, they are used for smoothing which is a particularly important issue in the context of small training corpus.

References

Lalit R. Bahl and Robert L. Mercer. 1976. Part-of-speech assignement by a statistical decision algorithm. *IEEE International Symposium on Information Theory*, pages 88–89.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Third Conference on Applied Computational Linguistics*, Trento, Italy.

Eric Brill. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. In *2nd Workshop on large Corpora*, Boston, USA.

Jean-Pierre Chanod and Pasi Tapanainen. 1995. Tagging French – comparing a statistical and a constraint-based method. In *EACL SIGDAT Workshop*, Dublin, Ireland. Association for Computational Linguistics - European Chapter.

Kenneth W. Church. 1989. A stochastic parts program noun phrase parser for unrestricted text. In *IEEE Proceedings of the ICASSP*, pages 695–698, Glasgow.

Kenneth W. Church. 1992. Current practice in part of speech tagging and suggestions for the future. In Simmons, editor, *Abornik praci: In Honor of Henry Kučera*. Michigan Slavic Studies.

Doug Cutting, Julian Kupiec, Jan Peterson, and Penelope Sibun. 1992. A practical part-of-speech tagger. Trento, Italy. Proceedings of the Third Conference on Applied Natural Language Processing.

Stephen DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Anne-Marie Derouault and Bernard Merialdo. 1986. Natural language modeling for phoneme-to-text transcription. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, volume 8(6), pages 742–749.

W. Nelson Francis and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin Company, Boston, Massachusetts. with the assistance of Andrew W. Mackie.

Stig Johansson. 1980. The LOB Corpus of British English Tests: presentation and comments. *Association for Literary and Linguistic Computing*, 1:25–36.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Antilla. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, New York.

S. Klein and R. F. Simmons. 1963. A grammatical approach to grammatical tagging coding of English words. *JACM*, 10:334–347.

Julian Kupiec. 1992. Robust part-of-speech tagging using HMM's. *Computers, Speech, and Language*, 6(3):225–242.

Geoffrey Leech, Roger Garside, and Erik Atwell. 1983. Automatic grammatical tagging of the LOB corpus. *ICAME News*, 7:13–33.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Corpus of the European Community Initiative. 1989 and 1990. Le monde newspaper.

Fernando Pereira, Michael Riley, and Richard Sproat. 1994. Weighted rational transductions and their application to human language processing. In *ARPA Workshop on Human Language Technology*, pages 249–254. Advanced Research Projects Agency, March 8–11.

Evelyne Tzoukermann, Dragomir R. Radev, and William A. Gale. 1995. Combining linguistic knowledge and statistical learning in French part-of-speech tagging. In *EACL SIGDAT Workshop*, Dublin, Ireland. Association for Computational Linguistics - European Chapter.

Atro Voutilainen. 1993. NPtool, a detector of English noun phrases. Columbus, Ohio. Proceedings of the Workshop on very large corpora.