Applying Extrasentential Context To Maximum Entropy Based Tagging With A Large Semantic And Syntactic Tagset

Ezra Black and **Andrew Finch** and **Ruigiang Zhang**
ATR Interpreting Telecommunications Laboratories
2-2 Hikaridai Seika-cho, Soraku-gun
Kyoto, Japan 619-02
black@itl.atr.co.jp
finch@itl.atr.co.jp
rzhang@itl.atr.co.jp

Abstract

Experiments are presented which measure the perplexity reduction derived from incorporating into the predictive model utilised in a standard tag-n-gram part-of-speech tagger, contextual information from previous sentences of a document. The tagset employed is the roughly-3000-tag ATR General English Tagset, whose tags are both syntactic and semantic in nature. The kind of extrasentential information provided to the tagger is semantic, and consists in the occurrence or non-occurrence, within the past 6 sentences of the document being tagged, of words tagged with particular tags from the tagset, and of boolean combinations of such conditions. In some cases, these conditions are combined with the requirement that the word being tagged belong to a particular set of words thought most likely to benefit from the extrasentential information they are being conjoined with. The baseline model utilized is a maximum entropy-based tag-n-gram tagging model, embodying a standard tag-n-gram approach to tagging: i.e. constraints for tag trigrams, bigrams, and and the word-tag occurrence frequency of the specific word being tagged, form the basis of prediction. Added into to this baseline tagging model is the extrasentential semantic information just indicated. The performance of the tagging model with and without the added contextual knowledge is contrasted, training from the 850,000-word ATR General English Treebank, and testing on the accompanying 53,000-word test treebank. Results are that a significant reduction in testset perplexity is achieved via the added semantic extrasentential information of the richer model. The model with both long-range tag triggers and more complex linguistic constraints achieved a perplexity reduction of 21.4%.

1 Introduction

It appears intuitively that information from earlier sentences in a document ought to help reduce uncertainty as to a word's correct part-of-speech tag. This is especially so for a large semantic and syntactic tagset such as the roughly-3000-tag ATR General English Tagset (Black et al., 1996; Black et al., 1998). And in fact, (Black et al., 1998) demonstrate a significant "tag trigger-pair" effect. That is, given that certain "triggering" tags have already occurred in a document, the probability of occurrence of specific "triggered" tags is raised significantly—with respect to the unigram tag probability model. Table 1, taken from (Black et al., 1998), provides examples of the tag trigger-pair effect.

Yet, it is one thing to show that extrasentential context yields a gain in information with respect to a unigram tag probability model. But it is another thing to demonstrate that extrasentential context supports an improvement in perplexity vis-a-vis a part-of-speech tagging model which employs state-of-the-art techniques: such as, for instance, the tagging model of a maximum entropy tag-n-gram-based tagger.

The present paper undertakes just such a demonstration. Both the model underlying a standard tag-n-gram-based tagger, and the same model augmented with extrasentential contextual information, are trained on the 850,000-word ATR General English Treebank (Black et al., 1996), and then tested on the accompanying 53,000-word test treebank. Performance differences are measured, with the result that semantic information from previous sentences within a document is shown to help significantly in improving the perplexity of tagging

#	Triggering Tag	Triggered Tag	I.e. Words Like:	Trigger Words Like:
1	NP1LOCNM	NP1STATENM	Hill, County, Bay	Utah, Maine, Alaska
2	JJSYSTEM	NP1ORG	national, federal	Party, Council
3	VVDINCHOATIVE	VVDPROCESSIVE	caused, died, made	began, happened
4	IIDESPITE	CFYET	despite	yet (conjunction)
5	DD	PPHO2	any, some, certain	them
6	PN1PERSON	LEBUT22	everyone, one	(not) only, (not) just
7	...	MPRICE	..., ..., ...	$452,983,000, $10,000
8	IIATSTANDIN	MPHONE22	at (sent.–final)	913-3434
9	IIFROMSTANDIN	MZIP	from (sent.–final)	22314-1698 (zip)
10	NNUNUM	NN1MONEY	25%, 12", 9.4m3	profit, price, cost

Table 1: Selected Tag Trigger–Pairs, ATR General–English Treebank

with the indicated tagset.

In what follows, Section 2 provides a basic overview of the tagging approach used (a maximum entropy tagging model employing constraints equivalent to those of the standard hidden Markov model). Section 3 discusses and offers examples of the sorts of extrasententially–based semantic constraints that were added to the basic tagging model. Section 4 describes the experiments we performed. Section 5 details our experimental results. Section 6 glances at projected future research, and concludes.

2 Tagging Model

2.1 ME Model

Our tagging model is a maximum entropy (ME) model of the following form:

$$p(t|h) = \gamma \prod_{k=0}^{K} \alpha_k^{f_k(h,t)} p_0 \qquad (1)$$

where:

- t is tag we are predicting;
- h is the history (all prior words and tags) of t;
- γ is a normalization coefficient that ensures: $\Sigma_{t=0}^{L} \gamma \prod_{k=0}^{K} \alpha_k^{f_k(h,t)} p_0 = 1$;
- L is the number of tags in our tag set;
- α_k is the weight of trigger f_k;
- f_k are trigger functions and $f_k \epsilon \{0, 1\}$;
- p_0 is the default tagging model (in our case, the uniform distribution, since all of the information in the model is specified using ME constraints).

The model we use is similar to that of (Ratnaparkhi, 1996). Our baseline model shares the following features with this tagging model; we will call this set of features the basic n-gram tagger constraints:

1. $w = X \;\&\; t = T$
2. $t_{-1} = X \;\&\; t = T$
3. $t_{-2} t_{-1} = XY \;\&\; t = T$

where:

- w is word whose tag we are predicting;
- t is tag we are predicting;
- t_{-1} is tag to the left of tag t;
- t_{-2} is tag to the left of tag t_{-1};

Our baseline model differs from Ratnaparkhi's in that it does not use any information about the occurrence of words in the history or their properties (other than in constraint 1). Our model exploits the same kind of tag–n–gram information that forms the core of many successful tagging models, for example, (Kupiec, 1992), (Merialdo, 1994), (Ratnaparkhi, 1996). We refer to this type of tagger as a tag–n–gram tagger.

2.2 Trigger selection

We use mutual information (MI) to select the most useful trigger pairs (for more details, see

(Rosenfeld, 1996)). That is, we use the following formula to gauge a feature's usefulness to the model:

$$\begin{aligned} MI(s,t) &= P(s,t)\log\frac{P(t|s)}{P(t)} \\ &+ P(s,\bar{t})\log\frac{P(\bar{t}|s)}{P(\bar{t})} \\ &+ P(\bar{s},t)\log\frac{P(t|\bar{s})}{P(t)} \\ &+ P(\bar{s},\bar{t})\log\frac{P(\bar{t}|\bar{s})}{P(\bar{t})} \end{aligned}$$

where:

- t is the tag we are predicting;
- s can be any kind of triggering feature.

For each of our trigger predictors, s is defined below:

Bigram and trigram triggers : s is the presence of a particular tag as the first tag in the bigram pair, or the presence of two particular tags (in a particular order) as the first two tags of a trigram triple. In this case, t is the presence of a particular tag in the final position in the n-gram.

Extrasentential tag triggers : s is the presence of a particular tag in the extrasentential history.

Question triggers : s is the boolean answer to a question.

This method has the advantage of finding good candidates quickly, and the disadvantage of ignoring any duplication of information in the features it selects. A more principled approach is to select features by actually adding them one-by-one into the ME model (Della Pietra et al., 1997); however, using this approach is very time-consuming and we decided on the MI approach for the sake of speed.

3 The Constraints

To understand what extrasentential semantic constraints were added to the base tagging model in the current experiments, one needs some familiarity with the ATR General English Tagset. For detailed presentations, see (Black et al., 1998; Black et al., 1996). An apercu can be gained, however, from Figure 1, which shows two sample sentences from the ATR Treebank (and originally from a Chinese take–out food flier), tagged with respect to the ATR General English Tagset. Each verb, noun, adjective and adverb in the ATR tagset includes a semantic label, chosen from 42 noun/adjective/adverb categories and 29 verb/verbal categories, some overlap existing between these category sets. Proper nouns, plus certain adjectives and certain numerical expressions, are further categorized via an additional 35 "proper–noun" categories. These semantic categories are intended for any "Standard–American–English" text, in any domain. Sample categories include: "physical.attribute" (nouns/adjectives/adverbs), "alter" (verbs/verbals), "interpersonal.act" (nouns/adjectives/adverbs/verbs/verbals), "orgname" (proper nouns), and "zipcode" (numericals). They were developed by the ATR grammarian and then proven and refined via day–in–day–out tagging for six months at ATR by two human "treebankers", then via four months of tagset–testing–only work at Lancaster University (UK) by five treebankers, with daily interactions among treebankers, and between the treebankers and the ATR grammarian. The semantic categorization is, of course, in addition to an extensive syntactic classification, involving some 165 basic syntactic tags.

Starting with a basic tag–n–gram tagger trained to tag raw text with respect to the ATR General English Tagset, then, we added constraints defined in terms of "tag families". A tag family is the set of all tags sharing a given semantic category. For instance, the tag family "MONEY" contains common nouns, proper nouns, adjectives, and adverbs, the semantic component of whose tags within the ATR General English Tagset, is "money": 500–stock, Deposit, TOLL–FREE, inexpensively, etc.

One class of constraints consisted of the presence, within the 6 sentences (from the same document)[1] preceding the current sentence, of one or more instances of a given tag family. This type of constraint came in two varieties: either including, or excluding, the words within the sentence of the word being tagged. Where these intrasentential words were included, they

[1] (Black et al., 1998) determined a 6-sentence window to be optimal for this task.

```
(_( Please_RRCONCESSIVE Mention_VVIVERBAL-ACT this_DD1 coupon_NN1DOCUMENT
when_CSWHEN ordering_VVGINTER-ACT

OR_CCOR ONE_MC1WORD FREE_JJMONEY FANTAIL_NN1ANIMAL SHRIMPS_NN1FOOD
```

Figure 1: Two ATR Treebank Sentences from Chinese Take–Out Food Flier (Tagged Only – i.e. Parses Not Displayed)

consisted of the set of words preceding the word being tagged, within its sentence.

A second class of constraints added to the requirements of the first class the representation, within the past 6 sentences, of related tag families. Boolean combinations of such events defined this group of constraints. An example is as follows: (a) an instance either of the tag family "person" or of the tag family "personal attribute"(or both) occurs within the 6 sentences preceding the current one; or else (b) an instance of the tag family "person" occurs in the current sentence, to the left of the word being tagged; or, finally, both (a) and (b) occur.

A third class of constraints had to do with the specific word being tagged. In particular, the word being classified is required to belong to a set of words which have been tagged at least once, in the training treebank, with some tag from a particular tag family; and which, further, always shared the same basic syntax in the training data. For instance, consider the words "currency" and "options". Not only have they both been tagged at least once in the training set with some member of the tag family "MONEY" (as well, it happens, as with tags from other tag families); but in addition they both occur in the training set only as nouns. Therefore these two words would occur on a list named "MONEY nouns", and when an instance of either of these words is being tagged, the constraint "MONEY nouns" is satisfied.

A fourth and final class of constraints combines the first or the second class, above, with the third class. E.g. it is both the case that some avatar of the tag family "MONEY" has occurred within the last 6 sentences to the left; and that the word being tagged satisfies the constraint "MONEY nouns". The advantage of this sort of composite constraint is that it is focused, and likely to be helpful when it does occur. The disadvantage is that it is unlikely to occur extremely often. On the other hand, constraints of the first, second, and third classes, above, are more likely to occur, but less focused and therefore less obviously helpful.

4 The Experiments

4.1 The Four Models

To evaluate the utility of long-range semantic context we performed four separate experiments. All of the models in the experiments include the basic ME tag–n–gram tagger constraints listed in section 2. The models used in our experiments are as follows:

(1) The first model is a model consisting ONLY of these basic ME tag–n–gram tagger constraints. This model represents the baseline model.

(2) The second model consists of the baseline model together with constraints representing extrasentential tag triggers. This experiment measures the effect of employing the triggers specified in (Black et al., 1998) —i.e. the presence (or absence) in the previous 6 sentences of each tag in the tagset, in turn— to assist a real tagger, as opposed to simply measuring their mutual information. In other words, we are measuring the contribution of this long-range information over and above a model which uses local tag–n–grams as context, rather than measuring the gain over a naive model which does not take context into account, as was the case with the mutual information experiments in (Black et al., 1998).

(3) The third model consists of the baseline model together with the four classes of more sophisticated question-based triggers defined in the previous section.

(4) The fourth model consists of the baseline model together with both the long-range

tag trigger constraints and the question-based trigger constraints.

We chose the model underlying a standard tag-n-gram tagger as the baseline because it represents a respectable tagging model which most readers will be familiar with. The ME framework was used to build the models since it provides a principled manner in which to integrate the diverse sources of information needed for these experiments.

4.2 Experimental Procedure

The performance of each the tagging models is measured on a 53,000-word test treebank hand-labelled to an accuracy of over 97% (Black et al., 1996; Black et al., 1998). We measure the model performance in terms of the perplexity of the tag being predicted. This measurement gives an indication of how useful the features we supply could be to an n-gram tagger when it consults its model to obtain a probablity distribution over the tagset for a particular word. Since our intention is to gauge the usefulness of long-range context, we measure the performance improvement with respect to correctly (very accurately) labelled context. We chose to do this to isolate the effect of the correct markup of the history on tagging performance (i.e. to measure the performance gain in the absence of noise from the tagging process itself). Earlier experiments using predicted tags in the history showed that at current levels of tagging accuracy for this tagset, these predicted tags yielded very little benefit to a tagging model. However, removing the noise from these tags showed clearly that improvement was possible from this information. As a consequence, we chose to investigate in the absence of noise, so that we could see the utility of exploiting the history when labelled with syntactic/semantic tags.

The resulting measure is an idealization of a component of a real tagging process, and is a measure of the usefulness of knowing the tags in the history. In order to make the comparisons between models fair, we use correctly-labelled history in the n-gram components of our models as well as for the long-range triggers. As a consequence of this, no search is nescessary.

The number of possible triggers is obviously very large and needs to be limited for reasons of practicability. The number of triggers used for these experiments is shown in Table 2. Using these limits we were able to build each model in around one week on a 600MHz DEC-alpha. The constraints were selected by mutual information. Thus, as an example, the 82425 question trigger constraints shown in Table 2 represent the 82425 question trigger constraints with the highest mutual information.

Description	Number
Tag set size	1837
Word vocabulary size	38138
Bigram trigger number	18520
Trigram trigger number	15660
Long history trigger number	15751
Question trigger number	82425

Table 2: Vocabulary sizes and number of triggers used

The improved iterative scaling technique (Della Pietra et al., 1997) was used to train the parameters in the ME model.

5 The Results

Table 4 shows the perplexity of each of the four models on the testset.

The maximum entropy framework adopted for these experiments virtually guarantees that models which utilize more information will perform as well as or better than models which do not include this extra information. Therefore, it comes as no surprise that all models improve upon the baseline model, since every model effectively includes the baseline model as a component.

However, despite promising results when measuring mutual information gain (Black et al., 1998), the baseline model combined only with extrasentential tag triggers reduced perplexity by just a modest 7.6% . The explanation for this is that the information these triggers provide is already present to some degree in the n-grams of the tagger and is therefore redundant.

In spite of this, when long-range information is captured using more sophisticated, linguistically meaningful questions generated by an expert grammarian (as in experiment 3), the perplexity reduction is a more substantial 19.4%.

#	Question Description	MI (bits)
1	Person or personal attribute word in full history	0.024410
2	Word being tagged has taken NN1PERSON in training set	0.024355
3	Person or personal attribute word in remote history	0.024294
4	Person or personal attribute or other related tags in full history	0.020777
5	Person or personal attribute or other related tags in remote history	0.020156

Table 3: The 5 triggers for tag NN1PERSON with the highest MI

#	Model	Perplexity	Perplexity Reduction
1	Baseline n-gram model	2.99	0.0%
2	Baseline + long-range tag triggers	2.76	7.6%
3	Baseline + question-based triggers	2.41	19.4%
4	Baseline + all triggers	2.35	21.4%

Table 4: Perplexity of the four models

The explanation for this lies in the fact that these question-based triggers are much more specific. The simple tag-based triggers will be active much more frequently and often inappropriately. The more sophisticated question-based triggers are less of a blunt instrument. As an example, constraints from the fourth class (described in the constraints section of this paper) are likely to only be active for words able to take the particular tag the constraint was designed to apply to. In effect, tuning the ME constraints has recovered much ground lost to the n–grams in the model.

The final experiment shows that using all the triggers reduces perplexity by 21.4%. This is a modest improvement over the results obtained in experiment 3. This suggests that even though this long-range trigger information is less useful, it is still providing some additional information to the more sophisticated question–based triggers.

Table 3 shows the five constraints with the highest mutual information for the tag NN1PERSON (singular common noun of person, e.g. lawyer, friend, niece). All five of these constraints happen to fall within the twenty-five constraints of any type with the highest mutual information with their predicted tags. Within Table 3, "full history" refers to the previous 6 sentences as well as the previous words in the current sentence, while "remote history" indicates only the previous 6 sentences. A "person word" is any word in the tag family "person", hence adjectives, adverbs, and both common and proper nouns of person. Similarly, a "personal attribute word" is any word in the tag family "personal attribute", e.g. left–wing, liberty, courageously.

6 Conclusion

Our main concern in this paper has been to show that extrasentential information can provide significant assistance to a real tagger. There has been almost no research done in this area, possibly due to the fact that, for small syntax–only tagsets, very accurate performance can be obtained labelling the Wall Street Journal corpus using only local context. In the experiments presented, we have used a much more detailed, semantic and syntactic tagset, on which the performance is much lower. Extrasentential semantic information is needed to disambiguate these tags. We have observed that the simple approach of only using the occurrence of tags in the history as features did not significantly improve performance. However, when more sophisticated questions are employed to mine this long-range contextual information, a more significant contribution to performance is made. This motivates further research toward finding more predictive features. Clearly, the work here has only scratched the surface in terms of the kinds of questions that it is possible to ask of the history. The maximum entropy approach that we have adopted is extremely accommodating in this respect. It is possible to go much further in the direction of querying the historical tag structure. For example, we can, in effect. exploit grammatical relations within previous sentences with an eye to predicting the tags of similarly related words in the current sentence. It is also possible to go even further and exploit the structure of full parses in the history.

References

E. Black, A. Finch, H. Kashioka. 1998. Trigger-Pair Predictors in Parsing and Tagging. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics, 17th Annual Conference on Computational Linguistics*, pages 131–137, Montreal.

E. Black, S. Eubank, H. Kashioka, J. Saia. 1998. Reinventing Part-of-Speech Tagging. *Journal of Natural Language Processing (Japan)*, 5:1.

E. Black, S. Eubank, H. Kashioka, R. Garside, G. Leech, and D. Magerman. 1996. Beyond skeleton parsing: producing a comprehensive large–scale general–English treebank with full grammatical analysis. In *Proceedings of the 16th Annual Conference on Computational Linguistics*, pages 107–112, Copenhagen.

S. Della Pietra, V. Della Pietra, J. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

J. Kupiec. 1992. Robust part-of-speech tagging using a hidden Markov model. In *Computer Speech and Language*, 6:225–242.

R. Lau, R. Rosenfeld, S. Roukos. 1993. Trigger-based language models: a maximum entropy approach. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, II:45–48.

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

A. Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech and Language*, 10:187–228.