Markov models for language-independent named entity recognition

Robert Malouf
Alfa-Informatica
Rijksuniversiteit Groningen
Postbus 716
9700AS Groningen
The Netherlands
malouf@let.rug.nl

1 Introduction

This report describes the application of Markov models to the problem of language-independent named entity recognition for the CoNLL-2002 shared task (Tjong Kim Sang, 2002).

We approach the problem of identifying named entities as a kind of probabilistic tagging: given a sequence of words $w_1 \ldots w_n$, we want to find the corresponding sequence of tags $t_1 \ldots t_n$, drawn from a vocabulary of possible tags T, which satisfies:

$$S = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 \ldots t_n | w_1 \ldots w_n) \quad (1)$$

The possible tags are `B-PER` and `I-PER`, which mark the beginning and continuation of personal names; `B-ORG` and `I-ORG`, which mark names of organizations; `B-LOC` and `I-LOC`, which mark names of locations; `B-MISC` and `I-MISC`, which mark miscellaneous names; and `O`, which marks non-name tokens.

We will assume that a sequence of tags can be modeled by Markov process, and that the probability of assigning a tag to a word depends only on a fixed context window (say, the previous word and tag). Thus, the sequence probability in (1) can be restated as the product of tag probabilities:

$$P(t_1 \ldots t_n | w_1 \ldots w_n) = \prod_{i=1,n} P(t_i | w_i, t_{i-1}, w_{i-1}, \ldots)$$

For each of the models described in the next section, the model parameters were estimated based on the provided training data, with no preprocessing or filtering. Then, the most likely tag sequence (based on the model) is selected for each sentence in the test data, and the results are evaluated using the `conlleval` script.

2 Models

In the first model (baseline), the tag probabilities depend only on the current word:

$$P(t_1 \ldots t_n | w_1 \ldots w_n) = \prod_{i=1,n} P(t_i | w_i)$$

The effect of this is that each word in the test data will be assigned the tag which occurred most frequently with that word in the training data.

The next model considered (HMM) is a simple Hidden Markov Model (DeRose, 1988; Charniak, 1993), in which the tag probabilities depend on the current word and the previous tag. Suppose we assume that the word/tag probabilities and the tag sequence probabilities are independent, or:

$$P(w_i | t_i, t_{i-1}) = P(w_i | t_i) P(t_i | t_{i-1}) \quad (2)$$

Then by Bayes' Theorem and the Markov property, we have:

$$\begin{aligned} P(t_1 \ldots t_n | w_1 \ldots w_n) &= \frac{P(w_1 \ldots w_n | t_1 \ldots t_n) P(t_1 \ldots t_n)}{P(w_1 \ldots w_n)} \\ &= \frac{\prod_{i=1,n} P(w_i | t_i) P(t_i | t_{i-1})}{P(w_1 \ldots w_n)} \end{aligned}$$

Since the probability of the word sequence $P(w_1 \ldots w_n)$ is the same for all candidate tag sequences, the optimal sequence of tags satisfies:

$$S = \underset{t_1 \ldots t_n}{\operatorname{argmax}} \prod_{i=1,n} P(w_i | t_i) P(t_i | t_{i-1}) \quad (3)$$

The probabilities $P(w_i|t_i)$ and $P(t_i|t_{i-1})$ can easily be estimated from training data. Using (3) to calculate the probability of a candidate tag sequence, the optimal sequence of tags can be found efficiently using dynamic programming (Viterbi, 1967).

While this kind of HMM is simple and easy to construct and apply, it has its limitations. For one,

(3) depends on the independence assumption in (2). In the next model (ME), we avoid this by using a conditional maximum entropy model to estimate tag probabilities. Maximum entropy models (Jaynes, 1957; Berger et al., 1996; Della Pietra et al., 1997) are a class of exponential models which require no unwarranted independence assumptions and have proven to be very successful in general for integrating information from disparate and possibly overlapping sources. In this model, the optimal tag sequence satisfies:

$$S = \operatorname*{argmax}_{t_1 \dots t_n} \prod_{i=1,n} P(t_i | w_i, t_{i-1})$$

where

$$P(t_i | w_i, t_{i-1}) = \frac{\exp\left(\sum_j \lambda_j f_j(t_{i-1}, w_i, t_i)\right)}{\sum_{\tau \in T} \exp\left(\sum_j \lambda_j f_j(t_{i-1}, w_i, \tau)\right)} \quad (4)$$

The indicator functions f_j 'fire' for particular combinations of contexts and tags. For instance, one such function might indicate the occurrence of the word *Javier* with the tag B-PER:

$$f(t_{i-1}, w_i, t_i) = \begin{cases} 1 & \text{if } w_i = \text{Javier } \& \ t_i = \texttt{B-PER} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

and another might indicate the tag sequence O B-PER:

$$f(t_{i-1}, w_i, t_i) = \begin{cases} 1 & \text{if } t_{i-1} = \texttt{O} \ \& \ t_i = \texttt{B-PER} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Each indicator f_j function also has an associated weight λ_j, which is chosen so that the probabilities (4) minimize the relative entropy between the empirical distribution $\tilde{P}$ (derived from the training data) and the model probabilities P, or, equivalently, which maximize the likelihood of the training data. Unlike the parameters of an HMM, there is no closed form expression for estimating the parameters of a maximum entropy model from the training data. So, we proceed iteratively, gradually refining the parameter estimates until the desired level of precision is reached. For these experiments, the parameters were fit to the training data using a limited memory variable metric algorithm (Malouf, in press).

The basic structure of the model is very similar to that of Borthwick (1999). However, in the models described here, no feature selection is performed. Also note that this formulation of maximum entropy Markov models differs slightly from that of McCallum et al. (2000). Here we use a single maximum entropy model, while McCallum, et al. use a separate model for each source state. Using separate models increases the sparseness of the training data and, at least for this task, slightly reduces the accuracy of the final tagger.

Using indicator functions of the type in (5) and (6), the model encodes exactly the same information as the HMM in (3), but with much weaker independence assumptions. This means we can add information to the model from partially redundant and overlapping sources. The model ME+ adds two additional types of information that were used by Borthwick (1999). It includes capitalization features, which indicate whether the current word is capitalized, all upper case, all lower case, mixed case, or non-alphanumeric, and whether or not the word is the first word in the sentence. And it also adds additional context sensitivity, so that the tag probabilities depend on the previous word, as well as the previous tag and the current word.

The next model, ME+m, adds one additional feature to ME+ that takes advantage of the structure of the training and test data. Often in newspaper articles, the first reference to an individual is by full name and title, while later references use only the person's surname. While an unfamiliar full name can often be identified as a name by the surrounding context, the surname appearing alone is more difficult to catch. For example, one article begins:

> El presidente electo de la Repblica Dominicana, *Hiplito Meja*, del Partido Revolucionario Dominicano (PRD) socialdemcrata, manifest que mantendr su apoyo a los XIV Juegos Panamericanos del 2003 en Santo Domingo. *Meja*, quien gan los comicios presidenciales en las votaciones del pasado 16 de mayo, asegur que ni l ni su partido cambiarn la posicin asumida ante el pueblo dominicano de respaldar la organizacin de los Juegos.

In the first sentence, the phrase *Hiplito Meja* can likely be identified as a personal name even if the surname is an unknown word, since the phrase consists of two capitalized words (the first a common first name) set off by commas. In the second sentence, however, *Meja* is much more difficult to identify as a name: a sentence-initial capitalized unknown word is most likely to be tagged as O. To allow the use in the first sentence to provide information about the second, ME+m uses a feature which is true just in case the current word occurred as part of a personal name previously in the text being tagged. With this feature, the model can take advantage of easy instances of names to help with more difficult instances later in the text.

All of the models described to this point are completely language independent and use no information not contained in the training data. The final model, ME+mf, includes one additional feature which indicates whether or not the current word appears in a list of 13,821 first names collected from a number of multi-lingual sources on the Internet. While the names are drawn from a wide range of languages and cultures, the emphasis is on European names, and in particular English and Spanish.

3 Results

Each of the models described in the previous section were trained using `esp.train` and evaluated on `esp.testa`. The results are summarized in Table 1.

As would be expected, HMM performs substantially better than baseline for every category but locations, though earlier cross-validation experiments suggest that this exception is an accident of the particular split between training and test data.

Perhaps more surprisingly, ME outperforms HMM by an even wider margin. In these two models, the tag probabilities are conditioned on exactly the same properties of the contexts. The only difference between the models is that the probabilities in ME are estimated in a way which avoids the independence assumption in (2). The poor performance of HMM suggests that this assumption is highly problematic.

Adding additional features, in ME+ and ME+m, offer further gains over the base model. However, the addition of a database of first names, in ME+mf, only slightly improves the performance on personal names and actually reduces the overall performance. This is likely due to the fact that the list of names contains many words which can also be used as locations and organizations. Perhaps the use of additional databases of geographic and non-personal names would help counteract this effect.

For the final results, the model which preformed the best on the evaluation data, ME+m, was trained on `esp.train` and evaluated with `esp.testa` and `esp.testb`, and trained on `ned.train` and evaluated with `ned.testa` and `ned.testb`. Before training, the part of speech tags were removed from

Method	Type	Precision	Recall	$F_{\beta=1}$
baseline	overall	44.59	43.52	44.05
	LOC	52.67	72.18	60.90
	MISC	22.27	22.52	22.40
	ORG	51.59	45.29	48.23
	PER	32.81	25.61	28.77
HMM	overall	44.03	42.97	43.50
	LOC	31.35	69.04	43.12
	MISC	44.09	25.23	32.09
	ORG	65.30	46.18	54.10
	PER	47.49	23.98	31.87
ME	overall	71.50	50.95	59.50
	LOC	66.36	72.49	69.29
	MISC	58.04	33.33	42.35
	ORG	73.67	49.26	59.04
	PER	81.80	42.31	55.77
ME+	overall	72.07	67.70	69.82
	LOC	63.84	77.26	69.91
	MISC	49.85	38.51	43.46
	ORG	77.45	59.45	67.27
	PER	80.48	82.00	81.23
ME+m	overall	74.78	71.07	72.88
	LOC	68.28	80.00	73.68
	MISC	56.51	37.16	44.84
	ORG	78.99	61.94	69.44
	PER	80.13	88.79	84.24
ME+mf	overall	74.55	70.45	72.44
	LOC	63.50	80.20	70.88
	MISC	54.63	38.51	45.18
	ORG	79.71	61.94	69.71
	PER	85.30	85.92	85.61

Table 1: Summary of preliminary models

`ned.train`, to allow a more direct cross-language comparison of the performance of ME+m.

The results of the final evaluation are given in Table 2. The performance of the model is roughly the same for both test samples of each language, though the performance differs somewhat between the two languages. In particular, the performance on `MISC` entities is quite a bit better for Dutch than it is for Spanish, and the performance on `PER` entities is quite a bit better for Spanish than it is for Dutch. These differences are somewhat surprising, as nothing in the model is language specific. Perhaps the discrepancy (especially for the `MISC` class) reflects differences in the way the training data was annotated; `MISC` is a highly heterogenous class, and the criteria for distinguishing between `MISC` and `O` entities is sometimes unclear.

4 Conclusion

The models described here are very simple and efficient, depend on no preprocessing or (with the exception of ME+mf) external databases, and yet provide a dramatic improvement over a baseline model. However, the performance is still quite a bit lower than results for industrial-strength language-specific named entity recognition systems.

There are a number of small improvements which could be made to these models, such as feature selection (to reduce overtraining) and the use of whole sentence sequence models, as in Lafferty et al. (2001) (to avoid the 'label-bias problem'). These refinements can be expected to offer a modest boost to the performance of the best model.

Acknowledgements

The research of Dr. Malouf has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences and by the NWO PIONIER project *Algorithms for Linguistic Processing.*

Spanish dev.	precision	recall	$F_{\beta=1}$
LOC	68.28%	80.00%	73.68
MISC	56.51%	37.16%	44.84
ORG	78.99%	61.94%	69.44
PER	80.13%	88.79%	84.24
overall	74.78%	71.07%	72.88

Spanish test	precision	recall	$F_{\beta=1}$
LOC	74.71%	70.57%	72.58
MISC	60.43%	40.88%	48.77
ORG	76.51%	74.43%	75.45
PER	72.63%	90.61%	80.63
overall	73.93%	73.39%	73.66

Dutch devel.	precision	recall	$F_{\beta=1}$
LOC	84.50%	58.40%	69.07
MISC	68.29%	60.32%	64.06
ORG	76.52%	42.71%	54.82
PER	54.55%	81.21%	65.27
overall	65.80%	61.06%	63.34

Dutch test	precision	recall	$F_{\beta=1}$
LOC	85.81%	68.22%	76.01
MISC	72.43%	59.98%	65.62
ORG	78.87%	47.66%	59.42
PER	61.03%	83.70%	70.59
overall	70.88%	65.50%	68.08

Table 2: Results obtained for the development and the test data sets for the two languages used in this shared task.

References

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22.

Andrew Borthwick. 1999. *A maximum entropy approach to named entity recognition*. Ph.D. thesis, New York University.

Eugene Charniak. 1993. *Statistical Language Learning*. MIT Press, Cambridge, MA.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:380–393.

Steven J. DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14:31–39.

E.T. Jaynes. 1957. Information theory and statistical mechanics. *Physical Review*, 106,108:620–630.

John Lafferty, Fernando Pereira, and Andrew McCallum. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Robert Malouf. in press. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Computational Language Learning (CoNLL-2002)*, Taipei.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 591–598.

Erik Tjong Kim Sang. 2002. CoNLL 2002 shared task. `http://lcg-www.uia.ac.be/conll2002/ner/`.

Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.