A Statistical Method for Handling Unknown Words

Alexander Franz
Computational Linguistics Program and Center for Machine Translation
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
amf@cs.cmu.edu

Robust Natural Language Processing systems must be able to handle words that are not in their lexicon. We created a classifier that was trained on tagged text to find the most likely parts of speech for unknown words. The classifier uses a contingency table to count the observed features, and a loglinear model to smooth the cell counts. After smoothing, the contingency table is used to obtain the conditional probability distribution for classification.

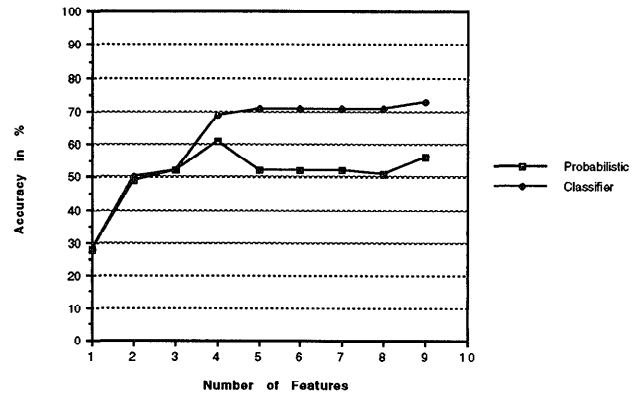
A number of features, determined by exploration (Tukey 1977), are used. For example, is the word capitalized? Does the word carry one of a number of known suffixes? We maximize the conditional probability of the proposed classification given the features to achieve minimum error rate classification (Duda & Hart 1973).

The baseline results are provided by using only the prior probabilities $P(c)$ (column **Prior**). (Weischedel *et al.* 1993) describe a probabilistic model with four features that are treated as independent, which we reimplemented (column **4 Indep**). For comparison, we created a statistical classifier with the same four features (column **4 Class**). Our best model was a classifier with nine features (column **9 Class**).

Measure	Prior	4 Indep	4 Class	9 Class
Overall Accuracy	28%	61%	69%	73%
Overall Res. Amb.	7.6	1.7	2.8	3.4
2-best Accuracy	53%	77%	87%	87%
2-best Res. Amb.	2.0	1.5	1.6	1.8
0.4-beam Accuracy		66%	81%	86%
0.4-beam Res. Amb.		1.2	1.4	1.6
0.4-beam Size		1.2	1.6	1.8

(n-best) Accuracy:	Percentage that the correct POS was among the n most likely POSs.
F-beam Accuracy:	All POSs with probability within beam factor F of the most probable POS.
Residual Ambiguity:	Mean perplexity for the POS tags in the answer set.
F-beam Size:	Mean number of tags in an answer set derived using beam factor F.

The graph below shows the accuracy of the simple probabilistic model versus the statistical classifier using one to nine features. The accuracy of the classifier is always higher and increases as more features are added, but does not decrease with nuisance features. The simple probabilistic model, on the other hand, peaks a four features, and then degrades.



In future work, we will apply this method to other ambiguity resolution problems that require a combination of a number of categorial disambiguating features, such as POS tagging and PP attachment.

Acknowledgments: I would like to thank Jaime Carbonell, Ted Gibson, Michael Mauldin, Teddy Seidenfeld, and Akira Ushioda.

References

Agresti, A. 1990. *Categorical Data Analysis.* New York: John Wiley & Sons.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* New York: John Wiley & Sons.

Franz, A. 1994. Ambiguity resolution via statistical classification: Clasifying unknown words by part of speech. Tecnical Report CMU-CMT-94-144, Center for Machine Translation, Carnegie Mellon University.

Tukey, J. 1977. *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Weischedel, R.; Meteer, M.; Schwartz, R.; Ramshaw, L.; and Palmucci, J. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics* 19(2):359–382.