Which -hunting and the Standard English relative clause: Online Supplement: Automatic *Zero* -Relative Detection

Joseph Frazee, Lars Hinrichs, Benedikt Szmrecsanyi, Axel Bohmann

Language, Volume 91, Number 4, December 2015, pp. s1-s3 (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.2015.0070*

WHICH-HUNTING AND THE STANDARD ENGLISH RELATIVE CLAUSE: ONLINE SUPPLEMENT: AUTOMATIC *ZERO*-RELATIVE DETECTION

JOSEPH FRAZEE
The University of Texas at Austin

LARS HINRICHS
The University of Texas at Austin

BENEDIKT SZMRECSANYI
KU Leuven

AXEL BOHMANN
The University of Texas at Austin

1. INTRODUCTION. It was nearly twenty years ago that computational linguists first argued for the use of statistical methods in order to make substantial contributions to areas of linguistics proper, including language variation and language change (Abney 1996). In that spirit we introduced a statistical supervised machine learning approach to the automatic detection of *zero*-relatives in the part-of-speech-tagged (POS-tagged) corpora. A brief overview of supervised machine learning is given in §2. Sections 3 and 4 describe the approach in more detail, and training and results are shown in §§5 and 6.

2. SUPERVISED MACHINE LEARNING. Machine learning, generally, is the subfield of computer science that is concerned with algorithms that have the capacity to 'learn' patterns based on task-specific criteria. Machine learning algorithms generate a model as the result of a period of training. The model, then, has the capacity to make decisions regarding the learned pattern. Machine learning has proven quite successful in application to natural language processing (NLP); and in statistical machine learning approaches to NLP, the training problem can be framed as either supervised or unsupervised learning.

Supervised machine learning algorithms are those that train a model to make decisions based on large numbers of observations from annotated corpora. For example, a supervised machine learning approach to POS tagging would produce a tagger by training a computer program on examples such as all of the word-tag pairs from the *Wall Street Journal* portion of the Penn Treebank (Marcus et al. 1994).

In contrast, unsupervised machine learning algorithms are not trained by exposure to labeled examples but instead by exposure to the unannotated examples themselves. At present, in NLP tasks, unsupervised machine learning models typically perform poorly as compared to supervised machine learners, making supervised approaches more advisable with the caveat that supervised training requires large amounts of annotated text.

Another dimension along which machine learners vary is whether they incorporate multiple perspectives on the data in the form of feature functions. Not all machine learners train on different views of training instances, but for those that do, the feature functions can be understood much like the feature systems of modern phonology and syntax. Each function f_i encodes a different opinion or value for each training token.

Keeping with the POS-tagging example, the feature functions would possibly include those that represent word form, whether the word ends in *-ed*, whether the word ends in *-tion*, and so on.

3. SYSTEM OVERVIEW AND DESCRIPTION. This supplement describes a supervised machine learning system trained for the task of identifying reduced relative clauses or zero-relatives in POS-tagged English-language corpora. The system employs a conditional random field framework (Lafferty et al.

2001) and is trained using features defined over *zero*-relative labelings from sections 02–18 of the Penn Treebank. The availability of tagged reduced relative clauses in the Penn Treebank is what makes this task feasible.

Given a stream of tagged text such as the one below:

(S1) The/DT man/NN I/PRP know/VB laughed/VBN ./.

the system produces a result as follows:

(S2) The/DT man/NN ZR/ZR I/PRP know/VB laughed/VBN ./.

The approach consists of three steps:

- Parsing and feature extraction from the Penn Treebank,
- Training, and
- Evaluation and application.

4. PARSING AND FEATURE EXTRACTION. The system uses features encoding word, POS, and capitalization (true/false) within a four-word window of every word boundary in a sentence. Long sequences of nonzeros (negative examples) are also pruned from the training set to avoid drowning out the contribution of the positive examples. In the above example, the second feature set would include features indicating that the preceding word is *man* and that the following word is *I*, that the preceding POS is NN and that the following POS is PRP, and so forth.

5. TRAINING. A conditional random field (CRF) was trained over sections 02–18 of the *Wall Street Journal* portion of the Penn Treebank. CRFs are conditionally trained, sequential models (Lafferty et al. 2001). They are conditionally trained, which means they can incorporate many independent features, as described above. This is in contrast to hidden Markov models (HMMs). CRFs, like HMMs, are, however, still sequence models. Sequence models are a class of algorithms that make each labeling decision with respect to each preceding and following decision. In the extant task, the CRF, then, uses the knowledge of adjacent features and labeling choices when deciding if a word boundary does in fact contain a zero-relative.

Accuracy	0.9994
Precision	0.6029
Recall	0.8541
F1	0.7068

TABLE S1. Precision, recall, F-measure, and accuracy for the *zero*-relative identification task.

	ACTUAL ZR	ACTUAL NOT ZR
PREDICTED ZR	123	81
PREDICTED NOT ZR	21	178,887

TABLE S2. Confusion matrix for the zero-relative identification task.

During training, a Gaussian prior with $\sigma^2 = 10.0$ was used to prevent overfitting; that is, the data was assumed to be drawn from a particular normal distribution in order to prevent the learner from learning to

label all instances negatively. After all, the number of word boundaries that do not contain zeros will certainly dwarf the number of word boundaries that do contain zeros.

6. RESULTS. In order to gauge the sort of performance to be expected in novel situations, the system was tested on sections 19–22 of the *Wall Street Journal* portion of the Penn Treebank. The results are presented in Tables S1 and S2 above. The confusion matrix in Table S2 suggests that for every 100,000 words containing ninety-six zero-relatives, the model will only fail to identify twelve zeros, and will mistakenly identify sixty-five nonzeros as zeros.

REFERENCES

ABNEY, STEVE. 1996. Statistical methods and linguistics. *The balancing act: Combining symbolic and statistical approaches to language*, ed. by Judith Klavans and Philip Resnik, 1–26. Cambridge, MA: MIT Press.

LAFFERTY, JOHN; ANDREW MCCALLUM; and FERNANDO C. N. PEREIRA. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, 282–89.

MARCUS, MITCHELL; BEATRICE SANTORINI; and MARY ANN MARCINKIEWICZ. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19.2.313–30. Online: https://www.aclweb.org/anthology/J/J93/J93-2004.pdf.

[joseph.frazee@gmail.com]
[larshinrichs@utexas.edu]
[benszm@kuleuven.be]
[axel.bohmann@utexas.edu]