Named Entity Recognition from Spontaneous Open-Domain Speech

Mihai Surdeanu, Jordi Turmo, Eli Comelles

Technical University of Catalunya
{surdeanu, turmo, comelles}@lsi.upc.edu

Abstract

This paper presents an analysis of named entity recognition and classification in spontaneous speech transcripts. We annotated a significant fraction of the Switchboard corpus with six named entity classes and investigated a battery of machine learning models that include lexical, syntactic, and semantic attributes. The best recognition and classification model obtains promising results, approaching within 5% a system evaluated on clean textual data.

1. Introduction

Named entity recognition and classification (NERC) is a fundamental component of many natural language processing (NLP) applications such as question answering, information extraction, clustering, topic detection, and summarization. While significant progress has been reported on the NERC task, most of the previous approaches have generally focused on clean textual data [6], or read speech data [3, 5], where most speech-specific phenomena are minimized.

On the other hand, in spontaneous speech a series of phenomena that make its processing difficult are emphasized: disfluency or stuttering, speaker corrections and specifications, and lack of grammatical structure. Additionally, spontaneous speech transcripts generally lack case information, a vital NERC clue in many languages. Nevertheless, spontaneous speech is a *de facto* attribute in many scenarios that could immediately benefit from the previously mentioned NLP applications: presentations, seminars, and meetings or other types of reunions or conversational acts.

In this paper we focus on the recognition and classification of named entities (NE) in spontaneous speech. We identify proper names (of persons, locations, organizations, and other categories classified under a miscellaneous label), temporal entities (dates and times of day), and monetary expressions. We propose a two-fold approach: first, we *annotate* a spontaneous speech corpus with the desired named entity categories, and second, using splits of this corpus for training and testing, we *investigate several machine learning models for NERC* that include various lexical, syntactic, and semantic features.

The contributions of this work are the following:

1. We analyze several machine learning models for NERC and identify which features are beneficial for the NERC task when applied to spontaneous speech transcripts.
2. We extend a spontaneous speech corpus with named entity annotations and show that the new data has a significant contribution to the NERC accuracy.

The paper is organized as follows: Section 2 introduces the corpus and the named entity categories used throughout the paper. Section 3 describes the recognition and classification models. Section 4 presents the experimental results and Section 5 concludes the paper.

```
B: You mean they don't have the, uh, the smog alerts?
A: No, not in, not in Te-, well not in Dallas, that is.
B: Right. I, I,
A: [throat_clearing].
B: yeah, I spent a summer i-, i-, in Tyler so I know,
   just east of Dallas there
A: Yeah. We're going there tomorrow.
```

Figure 1: *Sample Switchboard transcript fragment with two speakers: A and B.*

2. The Corpus

The corpus used in this paper is an extension of the Switchboard (SWB) corpus (LDC catalog number LDC97S62). SWB is a corpus of spontaneous telephone conversations containing over 240 hours of recorded speech and about 3 million words of text. Over 500 speakers of both sexes from every major dialect of American English were recorded. Figure 1 shows a sample fragment of a SWB transcript. Although the SWB transcripts contain case information, we will perform experiments both with and without this attribute.

Based on NE guidelines previously developed for the Message Understanding Conference (MUC) [2] and the Computational Natural Language Learning Conference (CoNLL) [6] we have annotated 6 classes of named entities in the SWB corpus:

1. Person names (*PER*) - this entity type includes not only real or fictional person names (”Stephen King”, ”Sue Ellen”), but also other fictional non-human individuals such as ”God”. Titles are not tagged, e.g. in the segment ”President Bush” we only tag ”Bush”.
2. Organization names (*ORG*) - proper names tagged as organizations include businesses, multinational organizations, stock exchanges, political parties, religious groups, governmental entity names, armies, etc. We also label metonymical locations when used as references to organizations, e.g. ”Washington” in ”Washington decided to invade Irak”.
3. Location names (*LOC*) - this type of classification applies to geographical, political or astronomical entities. We take into account proper nouns such as ”California”, as well as whole phrases such as ”southern California” when the constituents of these phrases specify the location.
4. Other names (*MISC*) - proper names that do not belong to PER, ORG, or LOC. This type of NEs include: titles of books, films and songs (”Bohemian Rapsody”); TV Shows (”Sesame Street”), names of fictitious animals (”Donald Duck”), events (”World War Two”), etc.
5. Temporal entities (*TIME*) - this category includes dates, such as ”March 28th”, or specific moments in time, such as ”last week”. We consider both absolute and relative time expressions. We include prepositions or other

phrase constituents when they help specify the time, e.g. "after two weeks".

6. Monetary entities (*MONEY*) - entities that indicate quantities of money. These NEs are generally composed of a quantity and a currency, such as "sixty thousand dollars", but appear also as the quantity alone, e.g. "sixty thousand".

The NE annotation process is still work in progress: we have currently annotated about 30% of the SWB data in approximately one person month. From the annotated documents we have selected 20% of the documents for testing and used the rest of the documents for training. Table 1 lists the number of entities of each class in the testing and training partitions.

To facilitate machine learning, we tokenize all transcripts and convert the NE annotations to the IOB2 format, which assigns to each token one of several labels: *B-CATEGORY* if the token begins a NE of type `CATEGORY`, *I-CATEGORY* if the token is inside a NE of type `CATEGORY`, and *O* if the token does not belong to any known entity [6].

3. NERC Models

One of the main purposes of the work presented in this paper is to investigate which attributes (be it lexical, syntactic, or semantic) are actually useful for the NERC task in spontaneous speech. To achieve this goal we propose a battery of 7 models, each one introducing a distinct set of attributes. The proposed models are cumulative: model M_n includes the attributes of all previous models, from M_1 to M_{n-1}.

All machine learning models proposed in this paper are implemented using Support Vector Machines (SVM) due to their capability to handle the large but sparse feature spaces (typical to NLP problems) with good generalization properties [4]. All SVMs were trained using one-vs-all classifiers with polynomial kernels of degree 2.

Model M_1 - contains only the following lexical attributes:

- The token lexem, both with and without case information (e.g. "IBM" and "ibm").
- The suffixes and prefixes of length 2, 3, and 4, for example "on", "son", and "nson" are the suffixes of the word "Johnson".
- The sequence obtained by removing all letters from the token, for example "&" for "AT&T".
- The sequence obtained by removing all alphanumeric characters from the token, for example "$--$" for "$10-06-2004$".

Model M_2 - adds the following format attributes:

- *isAllCaps* - Boolean flag set to true if the word contains only upper-case letters (e.g. "IBM").
- *isAllCapsOrDots* - Boolean flag set to true if the word contains only upper-case letter or dots (e.g. "I.B.M").
- *isAllDigits* - Boolean flag set to true if the word contains only digits.
- *isAllDigitsOrDots* - Boolean flag set to true if the word contains only digits or dots.
- *initialCap* - Boolean flag set to true if the word starts with an upper-case letter (e.g. "October").

	PER	ORG	LOC	MISC	TIME	MONEY
Train	1358	1444	3927	2837	1866	626
Test	313	328	982	747	482	101

Table 1: *Number of NEs of each class in the annotated corpus.*

Model M_3 - adds part of speech (POS) attributes. The POS tags are generated with a statistical POS tagger reported to have an accuracy of over 96% on several corpora [1].

Model M_4 - adds syntactic chunk labels. The syntactic chunks, i.e. simple non-recursive phrases such as nouns or verbs, are labeled using the same IOB2 format used for NEs. For example, the two tokens inside the noun phrase (NP) "Mr. Johnson" are labeled "B-NP" and "I-NP". The syntactic chunks are detected using the SVM-based framework reported best in the CoNLL shared task evaluation [4]. In our implementation, the chunker has an accuracy of over 95% on the CoNLL evaluation data.

Model M_5 - adds syntax-based context. We consider as context the right-most word, i.e. the *head word*, of the noun/verb syntactic chunks that precede/follow the current chunk. We skip context phrases whose head words are not nouns or verbs (e.g. pronouns). The intuition behind this model is that relevant nouns or verbs in the vicinity of the entity to be classified offer clues about its class. For example, the right context when classifying the entity "Ryan" in "...when Ryan struck out his five thousandth player they they..." is "struck" and "player", which offer strong hints that the name to be classified is a person's name.

Model M_6 - adds the following class-based attributes:

- *isNumber* - Boolean flag set to true if the token is a word-spelled number (e.g. "one", "nineteen").
- *isMultiplier* - Boolean flag set to true if the token is a multiplier typically used to compose numbers (e.g. "hundred", "thousand").
- *isDay* - Boolean flag set to true if the token is the name of a day of the week.
- *isMonth* - Boolean flag set to true if the token is the name of a month.

Model M_7 - adds gazetteer-based attributes. Using the same IOB2 notation, we indicate if a sequence of tokens is part of a known gazetteer. For example, the tokens in the sequence "Fort Collins" will have the attributes "B-LOC" and "I-LOC". For the experiments reported in this paper we have used four gazetteers: first and last person name gazetteers from the US Census (`http://www.census.gov`), a United States location gazetteer from USGS (`http://geonames.usgs.gov`), and a world location gazetteer from GeoNS (`http://earth-info.nga.mil`). The person name gazetteers contain over 90,000 entries and the location gazetteers contain over 5,000,000 known locations.

From the proposed models, M_1 and M_2 include only lexical attributes, M_3 to M_5 add morpho-syntactic attributes, and M_6 and M_7 add semantic features. Some of the proposed features (e.g. *isAllCaps* or *initialCap*) are case dependent, hence they are not used when training on data without case information. For robustness, the models M_6 and M_7 use case-insensitive search in the word lists or gazetteers.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7
LOC	81.13	83.50	83.92	83.34	82.71	82.18	83.11
MISC	56.14	65.57	65.39	65.24	65.48	66.49	66.36
MONEY	76.60	74.61	75.00	75.79	81.22	80.39	82.59
ORG	62.63	63.33	64.21	64.97	64.25	64.16	64.83
PER	62.48	74.27	72.73	70.76	69.22	70.30	77.72
TIME	76.68	75.50	75.90	76.00	76.24	75.88	75.37
Overall	**70.78**	**74.24**	**74.32**	**74.04**	**73.83**	**73.96**	**75.12**

Table 2: *NERC F measure on SWB transcripts* with *case information.*

	M_1	M_2	M_3	M_4	M_5	M_6	M_7
LOC	80.86	81.29	80.15	80.27	79.07	79.06	79.90
MISC	53.89	54.62	53.42	53.15	48.81	49.37	50.88
MONEY	78.31	78.72	77.49	76.84	79.59	79.00	79.41
ORG	61.79	62.68	62.68	63.19	58.71	59.85	58.10
PER	62.72	65.07	67.88	63.22	59.48	61.51	69.80
TIME	76.13	75.53	75.50	75.80	74.75	75.89	74.45
Overall	**70.09**	**70.56**	**70.09**	**69.67**	**67.55**	**68.13**	**69.22**

Table 3: *NERC F measure on SWB transcripts* without *case information.*

The SVM-based framework used in this paper has the capacity to extract attributes from a context window surrounding the current example. Static features (i.e. the attributes introduced by models M_1 to M_7) can be extracted both from the left and the right of the current example, while dynamic features (i.e. the classifier-assigned NE labels) are obviously considered only for the context elements previously labeled. Note that this context is different from the syntax-based context introduced by the model M_5. The latter may skip certain tokens that do not match the syntactic constraints, while the former can use only tokens in the immediate vicinity of the example to be labeled. Unless otherwise specified, all the proposed models used a context window of two neighboring tokens to the left and two to the right of the current example. The choice of this context window size is justified in the next section.

4. Experimental Results

4.1. Evaluation of The Proposed Models

In the first experiments we analyze the performance of the seven NERC models previously proposed. We have performed two sets of experiments. In the first set we used the original SWB transcripts, which (generally) contain case information. For the second set of experiments we converted all transcripts to lower case, thus discarding the case information. In both experiments we measure the F score (i.e. the harmonic mean of precision and recall) for each NE class and globally, for all NE classes combined. Table 2 summarizes the results for the data with case information. Table 3 lists the results for the SWB transcripts without case information.

The first observation is that the morpho-syntactic attributes (introduced in models M_3, M_4, and M_5) by and large do *not* help. POS tags (model M_3) induce a minor improvement in the experiment that uses data with case information, but have a negative effect when case is not available. The other syntax-based attributes, chunk labels and syntactic context, do not help at all. The explanation for this behavior is that the tools used to extract morpho-syntactic information (POS tagger and chunk detector) are trained on "clean" data with case information. When applied on the noisy spontaneous speech transcripts their performance is not sufficient to induce useful information. We expect this behavior to be even more apparent on transcripts generated by an automated speech recognizer (ASR).

CoNLL	SWB with case	SWB without case
80.52	75.50	71.55

Table 4: *Overall F measure of the best models on several corpora.*

On the other hand, the remaining lexical and semantic attributes all have an overall positive contribution. Word classes (model M_6) help most by eliminating spurious classifications as PER or MISC. As expected, gazetteer information (model M_7) helps the classification of entities in the PER and LOC classes. The gazetteer contribution is somewhat limited (especially for the LOC class) due to the ambiguity of the gazetteer data. For example, common English words like "How" and "To" appear as Asian location names in the world gazetteer.

The above observations indicate that the NERC task on spontaneous speech is another instance of the "less is more" situation: the best NERC model should include only lexical (models M_1, M_2) and semantic (models M_6, M_7) attributes. If the data has case information, the attribute set should also include POS tags. Following these guidelines, we have constructed two models: M_c, which includes attributes introduced in models M_1, M_2, M_3, M_6, and M_7 and is tailored for data *with case* information, and M_{nc}, which includes attributes from M_1, M_2, M_6, and M_7 and is tailored for data *without case* information.

Table 4 summarizes the performance of these two best models on three corpora: (a) clean textual data from the CoNLL shared task evaluation [6], (b) SWB transcripts with case, and (c) SWB transcripts without case. For this comparative analysis we have selected a subset of the CoNLL training data that has the same number of positive training examples as SWB (about 40% of the complete CoNLL corpus). Table 4 indicates that the NERC performance on spontaneous speech is very promising. The speech-specific phenomena account for a F measure drop of 5% from the F measure of the system trained and tested on textual data. The elimination of case accounts for another 4% drop. Nevertheless, considering that we have currently annotated only 30% of the SWB corpus, we see a performance of over 70% F measure on spontaneous speech transcripts without case information as very encouraging.

4.2. Contribution of Context

All models evaluated in the previous sub-section used an immediate context window of 2 tokens to the left and 2 tokens to the right of the current token. Table 5, which analyzes the behavior of the M_{nc} model for various context windows, justifies this choice. Table 5 shows that MONEY and TIME generally benefit from larger context windows, while LOC, MISC, ORG, and PER do not. MONEY and TIME gain from larger contexts mainly because entities in these classes include more tokens than the other entities, hence a larger window is needed to reach the actual context surrounding the entity. The best compromise is achieved for a context window of 2 elements to the left and 2 to the right of the token to be labeled (± 2). This result is an empirical proof that a fairly small context is sufficient for NERC.

	±1	±2	±3	±4
LOC	80.18	82.34	80.35	79.19
MISC	55.15	54.55	50.25	46.17
MONEY	69.16	79.38	80.40	82.05
ORG	64.83	64.37	60.68	58.72
PER	73.58	71.45	69.01	66.78
TIME	69.20	74.13	77.20	74.95
Overall	**69.79**	**71.55**	**69.91**	**67.95**

Table 5: *Contribution of context to the system performance using the M_{nc} model.*

LOC	MISC	ORG	PER	**Overall**
-13.12	-43.45	-30.66	-22.19	**-24.68**

Table 6: *F measure drop when training the NERC model on "clean" textual data and testing on SWB.*

4.3. Justification for The Spontaneous-Speech Corpus

A legitimate question about the work presented in this paper is why is yet another NE corpus useful? We argued that spontaneous speech poses a unique challenge and the best way to tackle it is to annotate a dedicated corpus. We prove that textual data does not capture the spontaneous speech characteristics by training the model M_c on the CoNLL corpus and testing it on SWB. Table 6 shows the drop in F measure from the results obtained when training M_c on the actual SWB transcripts. For this experiment we used only the NE classes that were annotated using the same specification in the two corpora. The overall drop of almost 25 absolute percents indicates that, indeed, textual data is far from modeling the spontaneous-speech phenomena.

4.4. Effect of Training Set Size

To assess the effect of training set size on NERC in spontaneous speech, we trained the M_{nc} model on successively larger subsets of the SWB training partition, testing every time on the complete testing set. Figure 2 summarizes the results. While the NERC F measure is over 60% with as little as 25% of the current training partition, performance continues to increase as the size of the training set increases. This is strong motivation to continue our annotation work in the SWB corpus. Since we have currently annotated only one third of the SWB corpus, when the whole corpus is annotated we expect the final F measure to be in the 80 percents, an excellent performance even for clean textual data [6].

5. Conclusions

In this paper we focus on the recognition and classification of named entities in spontaneous speech. We identify proper names (of persons, locations, organizations, and other categories classified under a miscellaneous label), temporal entities (dates and times of day), and monetary expressions.

To achieve state-of-the-art performance we propose a two-step approach: first, we annotate a spontaneous speech corpus with the above NE categories, and second, using this corpus we investigate the performance of a battery of machine learning models that include lexical, syntactic, and semantic attributes. We conclude that the best model includes lexical and semantic attributes. Syntactic attributes by and large do not help, with the exception of part-of-speech tags which have a minor contribution.

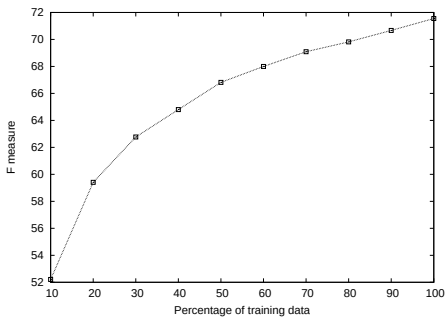

Figure 2: *Effect of training set size on the overall F measure.*

The overall performance on spontaneous speech is very encouraging, approaching within 5% the F measure of the system trained and tested on clean textual data. If case information is not available in the speech transcripts, the system performance is still within 9% of the system evaluated on textual data. These results indicate that speech-specific phenomena and lack of case information do affect NERC performance, but the task can be tackled with promising results.

Furthermore, we show that the system performance continues to increase as more training data becomes available, which motivates us to continue our annotation work on the spontaneous speech corpus. Since we have currently annotated only a third of the speech corpus, we estimate that the final system performance will reach or surpass 80% when the whole corpus is annotated.

6. Acknowledgments

This work has been partially funded by the European project CHIL (IP-506808) and the Spanish Ministry of Science and Technology project TIN2004-0171-E. Mihai Surdeanu is a research fellow within the Ramón y Cajal program of this ministry.

7. References

[1] Brants, T., "TnT – A Statistical Part-of-Speech Tagger", In Proc. of the 6th Applied NLP Conference, 2000.

[2] Chinchor, N., Brown E., Ferro, L., and Robinson, P., "1999 Named Entity Recognition Task Definition", MITRE and SAIC, 1999.

[3] Kubala, F., Schwartz, R., Stone, R., and Weischedel, R., "Named Entity Extraction from Speech", In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[4] Kudo, T. and Matsumoto, Y., " Fast Methods for Kernel-Based Text Analysis", In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), 2003.

[5] Palmer, D.D., Burger, J.D., Ostendorf, M., "Information Extraction from Broadcast News Speech Data", In Proc. of the DARPA Broadcast News Workshop, 1999.

[6] Sang, E.F.T.K. and De Meulder, F., "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition", In Proc. of CoNLL, 2003.