

# A Simple Yet Effective Approach for Named Entity Recognition from Transcribed Broadcast News

Md. Faisal Mahbub Chowdhury

HLT Research Unit, FBK-Irst, Trento, Italy  
Department of Information Eng. and Computer Science, University of Trento, Italy  
chowdhury@fbk.eu

**Abstract.** Automatic speech transcriptions pose serious challenges for NLP systems due to various peculiarities in the data. In this paper, we propose a simple approach for NER on speech transcriptions which achieves good result despite the peculiarities. The novelty of our approach is that it emphasizes on the maximum exploitation of the tokens, as they are, in the data. We developed a system for participating in the “NER on Transcribed Broadcast News” (closed) task of the EVALITA 2011 evaluation campaign where it was the 2nd best system out of 4 participant teams with an F1-score of *57.02* on the automatic speech transcription test data. On the manual transcriptions of the same test data (although having no sentence boundary and punctuation symbol), the system achieves an F1-score of *73.54* which is quite high considering the fact that the system is language independent and uses no external dictionaries, gazetteers or ontologies.

**Keywords:** Named entity recognition, automatic speech transcription, transcribed broadcast news.

## 1 Introduction

Named Entity Recognition (NER) on written texts achieved significant progress in recent years. This has contributed in the growing interest for NER on speech transcripts. However, automatic speech transcriptions pose serious challenges for natural language processing (NLP) systems due to errors in word recognition, capitalization and segmentation as well as missing sentence boundaries and punctuation symbols. As a consequence, a system designed to perform specific tasks such as NER on such data is required to be resilient enough to overcome these challenges. Earlier evaluation campaigns for NER on broadcast news transcripts which include the DARPA Broadcast News Transcription and Understanding Workshop<sup>1</sup>, the Ester 2 evaluation campaign<sup>2</sup>, etc envisaged to advance the state of the art where participants used complex methodologies.

<sup>1</sup> <http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa98/index.htm>

<sup>2</sup> [http://www.afcp-parole.org/camp\\_eval\\_systemes\\_transcription/](http://www.afcp-parole.org/camp_eval_systemes_transcription/)

We developed a system as part of our participation in the “NER on Transcribed Broadcast News” (closed) task of the EVALITA 2011 evaluation campaign where it was the 2nd best system out of 4 participants with an F1-score of *57.02* on the automatically transcribed broadcast news. On the manual transcriptions of the same test data, where word errors are fixed but sentence boundaries and punctuation symbols are still missing, the system achieves an F1-score of *73.54* which is quite good considering the simple approach of our system.

The simplicity and novelty of our approach which differ from previous work lie in the fact that we do not use any external resources<sup>3</sup> or complex NLP toolkit. We even do not try to recover capitalization, digits, punctuation symbols and sentence boundaries (that are missing inside automatic speech transcriptions) using sophisticated linguistic methods and resources. Our approach emphasizes on the maximum exploitation of the tokens themselves and their distribution as they are in the data. Moreover, our approach is language independent.

The remainder of the paper is organized as follows. In Section 2, we briefly review the EVALITA 2011 NER task and data. Following that, Section 3 describes our proposed approach and how it is implemented. Finally, empirical results are discussed in Section 4.

## 2 EVALITA 2011 NER (closed) Task

The “NER on Transcribed Broadcast News” task<sup>4</sup> of the EVALITA 2011 evaluation campaign has two modalities – ‘*open*’ and ‘*closed*’. Due to time limitation, we only participated on the ‘*closed*’ modality which was compulsory. In this modality *only the training data distributed by the organizers and no additional resources were allowed* for training and tuning the system. Complex NLP toolkits which use named entity (NE) dictionaries were also forbidden for the closed run, while simple tools for POS tagging or lemmatization were allowed. Participants were expected to identify four NE types: (i) *Person (PER)*, (ii) *Organization (ORG)*, (iii) *Location (LOC)*, and (iv) *Geo-Political Entities (GPE)*.

The test data consist of unannotated recorded and automatically transcribed broadcast news of 10 audio files. The training data consist of the following:

- the RTTR data
  - 10 broadcast news program manually transcribed and annotated with NEs.
  - both automatic transcriptions and audio files of the same news.
- the I-CAB data [2], a corpus of (written) news stories annotated with Named Entities including the data used for EVALITA 2007 and 2009 NER tasks.

The RTTR training data do not have have any POS (parts-of-speech) tags or lemmatization information. Instead, each line of the data contain individual

---

<sup>3</sup> For example, gazetteers, NE dictionaries, ontologies, Wikipedia or any kind of external lexicon or list.

<sup>4</sup> <http://www.evalita.it/2011/tasks/NER>

token, corresponding news program id, and an IOB2 tag indicating whether the token is part of an entity or not. The test data have similar formatting as the RTTR training data except that there is no IOB2 annotation.

There is no sentence boundary and punctuation symbol (e.g. comma (,)) in both manual and automatic transcriptions of the RTTR training and test data<sup>5</sup>. Moreover, the automatic transcriptions of the RTTR training and test data contain transcription errors (both in terms of word recognition and segmentation, and in terms of word capitalization).

The I-CAB data contain POS tag annotations. Although we exploit the I-CAB data (more specifically, the EVALITA 2007 and 2009 NER data), we ignore the POS tag annotations.

### 3 Description of Our System

Our system has two components. The first component is responsible for automatically building dictionaries from the training data as well as for lemmatization and POS tagging of the training and test data. The second component collects features from the training data and dictionaries (processed by the first component), trains a first-order conditional random field (CRF) model, and annotates the test data.

For CRF model training, the Mallet toolkit [3] has been used. The lemmatization and POS tagging has been done using the TreeTagger tool [4]. Both the training (i.e. the I-CAB and RTTR data) and test data have been lemmatized and POS tagged.

At first, our system automatically builds the following dictionaries from the training data:

- *Dictionary of tokens with same POS tag (DictPOS)* : List of those tokens (and their POS tags) whose corresponding POS tags are always the same in the I-CAB data.
- *Dictionary of non-entity tokens (DictNonEnt)* : List of tokens which are always labelled as “O” in the I-CAB and RTTR training data.
- *Dictionary of entities (DictEntity)* : List of token sequences (along with corresponding NE types) which are always labelled as one of the NE types (i.e. *LOC*, *PER*, *ORG*, and *GPE*) in the I-CAB and RTTR training data.

As the descriptions of the dictionaries imply, our system filters any token from being included in *DictPOS* and *DictNonEnt* whose POS tag and IOB2 label (as “O”) respectively vary in the corresponding training data used. Likewise, if there is a token sequence which is annotated as an NE in somewhere in the training data and also annotated as not NE (i.e. “O”) somewhere else, the system does not consider it for *DictEntity*. For example, in the I-CAB data, “scuola” is annotated as both *ORG* and *other* (i.e. “O”). The entities inside *DictEntity* are sorted in decreasing order according to their number of tokens.

---

<sup>5</sup> Note, manual transcriptions of the test data were released after the challenge.

Previously, [1] showed that POS tagging highly depends on adjacent tokens. The RTTR training and test data have no punctuation symbol and sentence boundary inside it. So, it is very likely that the POS tag annotation inside them would be more susceptible to errors. The objective behind the creation of *DictPOS* is to reduce such errors in these datasets. Our system looks for any token in the RTTR training and test data which also exists in *DictPOS*, and then change its POS tag in the data according to its tag in *DictPOS*.

The following feature types are extracted for training a CRF model:

- **General features:**
  - *Token*: The original token itself.
  - *Lemma*: Lemmatized form of the token.
  - *POS*: Part-of-speech tag of the token.
  - *charNgram*: 3 and 4 character n-grams.
  - *Suffix*: 2–4 character suffixes.
- **Contextual features:**
  - Bi-grams of  $\text{Token}_{k,k+1}$  where  $i - 2 \leq k < i + 2$ .
  - Bi-grams of  $\text{POS}_{k,k+1}$  where  $i - 2 \leq k < i + 2$ .
  - Bi-grams of  $\text{Lemma}_{k,k+1}$  where  $i - 2 \leq k < i + 2$ .
  - Offset conjunctions extracted by Mallet from features in the range from  $\text{token}_{i-1}$  to  $\text{token}_i$ .
- **Orthographic features:**
  - *InitCap*: whether initial letter is capital.
  - *AllCap*: whether all letters are capitals.
  - *SingLow*: whether the token is a single lower case letter.
  - *SingUp*: whether the token is a single upper case letter.
  - *HasEndingI*: whether the last character of the token is ‘i’.
  - *HasEndingO*: whether the last character of the token is ‘O’.
- **Dictionary lookup features**<sup>6</sup>
  - *TaggedOther*: whether the token is found in *DictNonEnt*.
  - *FoundInORG*: whether the token is part of an entity name of NE type *ORG* in *DictEntity*.
  - *FoundInPER*: whether the token is part of an entity name of NE type *PER* in *DictEntity*.
  - *FoundInLOC*: whether the token is part of an entity name of NE type *LOC* in *DictEntity*.
  - *FoundInGPE*: whether the token is part of an entity name of NE type *GPE* in *DictEntity*.

---

<sup>6</sup> During dictionary creation, any character other than *a-z, A-Z’, à, è, é, ì, í, î, ò, ó, ù, ú*, and *0-9* in the tokens of the dictionaries is removed. The same changes are applied to the tokens of the training and test data before matching them with the dictionary entries.

The orthographic features *HasEndingO* and *HasEndingI* are used since most of the Italian first and last names end with ‘*O*’ and ‘*i*’ correspondingly. However, we observed that they have a very minor impact on the results.

Once training is done, the system annotates the test data and then apply post-processing techniques. Post-processing is done for both tokens tagged as “*O*” (i.e. not part of any NE) and the token sequences tagged as an NE. If there is a token sequence which is annotated as an NE, then we look for it inside *DictEntity*. If a match is found, the NE type of the token sequence in the dictionary is matched with its NE type in the test data. If they are not same, the NE type inside the test data is changed according to that inside the dictionary.

For any token tagged as “*O*” inside the test data, the system looks for all the entity names which contain the token. For each of those entities, the system searches for the longest match, in terms of number of tokens, surrounding and including the token inside the test data. In case of a match, if all of the corresponding tokens of the match inside the test data are tagged as “*O*”, their IOB2 tags are changed accordingly to the tokens of the corresponding entity name.

## 4 Results and Discussion

**Table 1.** Results (provided by the organizers) on the test data of the EVALITA 2011 NER (closed) task. Accuracy obtained: *95.89%*.

NE type	Total NEs found	Correct NEs	Precision	Recall	F1-score
GPE	660	-	75.45	73.78	74.61
LOC	55	-	69.09	40.43	51.01
ORG	371	-	52.02	35.41	42.14
PER	364	-	51.92	41.45	46.10
ALL	1450	918	63.31	51.86	57.02

**Table 2.** Results (provided by the organizers) on the manual transcriptions of the test data of the EVALITA 2011 NER (closed) task. Accuracy obtained: *97.35%*.

NE type	Total NEs found	Correct NEs	Precision	Recall	F1-score
GPE	650	-	84.15	81.40	82.75
LOC	58	-	84.48	55.68	67.12
ORG	468	-	63.89	56.74	60.10
PER	368	-	81.25	71.88	76.28
ALL	1544	1194	77.33	70.11	73.54

Table 1 shows our result provided by the task organizers on EVALITA 2011 NER (closed) task. As we can see, the overall results (F1-score 57.02) is fair enough considering the complexity of the task. When the word errors are fixed, the results reach up to 73.54 (see Table 2).

Most noticeably, there is a huge increase in the identification of total number of *ORG*. While all of the performance indicators significantly improved for *PER*, *LOC*, and *GPE*, their total number remains relatively the same. This indicates that the improvements might be primarily gained due to the decrement of the boundary disagreements between the gold annotations and the system annotations. Nevertheless, it is evident that word errors play a significant role in correct identification. Specifically, it seems from the results that PERSON names are most affected by word errors.

To conclude, we have presented a simple language independent system for NER on automatic speech transcriptions. The system can be further improved by using external components to recover capitalization, digits, punctuation symbols and sentence boundaries in the speech transcriptions, and also by utilizing resources such as gazetteers or NE dictionaries.

## Acknowledgments

The author would like to thank all the task organisers (especially Manuela Speranza) and also Alberto Lavelli for clarifications regarding the task description and requirements.

## References

1. Chowdhury, M.F.M., Negri, M.: Expected answer type identification from unprocessed noisy questions. In: Andreasen, T., Yager, R., Bulskov, H., Christiansen, H., Larsen, H. (eds.) Flexible Query Answering Systems (FQAS '09), LNCS, vol. 5822, pp. 263–274. Springer, Heidelberg (2009)
2. Magnini, B., Pianta, E., Speranza, M., Lenzi, V.B., Sprugnoli, R.: Italian content annotation bank (I-CAB): Named entities. Technical report, FBK (2011)
3. McCallum, A.K.: Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu> (2002)
4. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing, pp. 44–49. Manchester, UK (1994)