Named Entity Recognition in Broadcast News Using Similar Written Texts

Niraj Shrestha
KU Leuven, Belgium
niraj.shrestha@cs.kuleuven.be

Ivan Vulić
KU Leuven, Belgium
ivan.vulic@cs.kuleuven.be

Abstract

We propose a new approach to improving named entity recognition (NER) in broadcast news speech data. The approach proceeds in two key steps: (1) we detect block alignments between highly similar blocks of the speech data and corresponding written news data that are easily obtainable from the Web, (2) we employ term expansion techniques commonly used in information retrieval to recover named entities that were initially missed by the speech transcriber. We show that our method is able to find the named entities missing in the transcribed speech data, but also to correct incorrectly assigned named entity tags. Consequently, our novel approach improves state-of-the-art results of NER from speech data both in terms of recall and precision.

1 Introduction

Named entity recognition (NER) is a task of extracting and classifying information units like *persons*, *locations*, *time*, *dates*, *organization names*, etc (e.g., Nadeau and Sekine (2007)). In general, the task involves labeling (proper) nouns with suitable *named entity tags*. NER is a very important pre-processing task in many applications in the fields of information retrieval (IR) and natural language processing (NLP). NER from speech data also displays its utility in various multimedia applications. For instance, it could be used in indexing video broadcast news using the associated speech data, that is, assigning names and their semantic classes recognized from the speech data as metadata to the video sequences (Basili et al., 2005).

NER from speech data is a difficult task and current state-of-the-art results are typically much lower than the results obtained from written text. For instance, the Stanford NER system in the CoNLL 2003 shared task on NER in written data report an F_1 value of 87.94% (Stanford, 2003). (Kubala et al., 1998; Miller et al., 1999) report a degrade of NER performance between 20-25% in F_1 value when applying a NER trained on written data to transcribed speech.

This lower performance has several causes. Firstly, speech transcribers often incorrectly transcribe phrases and even complete sentences, which might consequently result in many missing named entities. Secondly, many names were typically not observed in the training data on which the speech transcriber was trained (e.g., the problem is especially prominent when dealing with dynamic and ever-changing news data). The transcription then results in names and surrounding context words that are wrongly spelled, making the named entity recognition even more challenging. Finally, the named entity recognizer, especially when dealing with such unseen words, might incorrectly recognize and classify the named entities, and even tag non-names with named entity tags.

In this paper, we focus on the first two problems. We assume that similar written texts discussing the same news events provide additional knowledge about the named entities that are expected to occur in the spoken text. This external knowledge coming from written data then allows finding missing names and correcting incorrectly assigned named entity tags.

We utilize *term expansion and pseudo-relevance feedback techniques* often used in IR. The general idea there is to enrich queries with related terms. These terms are extracted from documents that were selected as being relevant for the query by the user or automatically by the IR system (Cao et al., 2008). Only certain terms are selected for expansion based on their importance in the relevant document and their

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 142–148,
Hissar, Bulgaria, 9-11 September 2013.

semantic relation with the query. We apply a similar approach to expanding and correcting the set of named entities in a speech document by the named entities found in the related relevant written documents. Following this modeling intuition, we are able to improve the recall of NER from broadcast speech data by almost 8%, while precision scores increase for around 1% compared to the results of applying the same named entity recognizer on the speech data directly. The contributions of this article are:

- We show that NER from speech data benefits from aligning broadcast news data with similar written news data.
- We present several new methods to recover named entities from speech data by using the external knowledge from high-quality similar written texts.
- We improve the performance of the state-of-the-art Stanford NER system when applied to the transcribed speech data.

The following sections first review related research, describe the methodology of our approach and the experimental setup, and finally present our evaluation and discuss the results.

2 Related Work

Named entity recognition was initially defined in the framework of Message Understanding Conferences (MUC) (Sundheim, 1995a). Since then, many conferences and workshops such as the following MUC editions (Chinchor, 1997; Sundheim, 1995a), the 1999 DARPA broadcast news workshop (Przybocki et al., 1999) and CoNLL shared tasks (Sang, 2002; Sang and Meulder, 2003) focused on extending the state-of-the-art research on NER. One of the first NER systems was designed by Rau (1991). Her system extracts and identifies company names by using hand-crafted heuristic rules. Today, NER in written text still remains a popular task. State-of-the-art NER models typically rely on machine learning algorithms trained on documens with manually annotated named entities. Examples of publicly available NER tools are the Stanford NER, OpenNLP NameFinder[1], Illinois NER system[2], the lingpipe NER system[3].

[1] http://opennlp.sourceforge.net/models-1.5
[2] http://cogcomp.cs.illinois.edu/page/software_view/4
[3] http://alias-i.com/lingpipe/web/models.html

NER in speech data poses a more difficult problem. In speech data and its transcribed variants, proper names are not capitalized and there are no punctuation marks, while these serve as the key source of evidence for NER in written data. Additionally, speech data might contain incorrectly transcribed words, misspelled words and missing words or chunks of text which makes the NER task even more complex (Sundheim, 1995b; Kubala et al., 1998).

NER in speech data was initiated by Kubala (1998). He applied the NER on transcription of broadcast news and reported that the performance of NER systems degraded linearly with the word error rate of the speech recognition (e.g., missing data, misspelled data and spuriously tagged names). Named entity recognition of speech was further investigated, but the relevant research typically focuses on improved error rates of the speech transcriptions (Miller et al., 1999; Palmer and Ostendorf, 2001), on considering different transcription hypotheses of the speech recognition (Horlock and King, 2003; Béchet et al., 2004) and on the problem of a temporal mismatch of the training data for the NER and the test data (Favre et al., 2005). None of these articles consider exploiting external text sources to improve the NER nor the problem of recovering missing named entities in the speech transcripts. .

3 Methodology

The task is to label a sequence of words $[w_1, w_2, \ldots, w_N]$ with a sequence of tags $[t_1, t_2, \ldots, t_N]$, where each word $w_i, i = 1, \ldots, N$ is assigned its corresponding tag $t_i \in \{person, organization, location\}$ in the transcribed speech of broadcast news.

3.1 Basic Architecture

The straightforward approach to NER in speech data is to apply the NER tagger such as Stanford NER tagger (Stanford, 2012) directly to transcribed speech data. However, the tagger will miss or assign incorrect named entity tags to many named entities due to the inherent errors in the transcription process. In this paper, we use related written text to recover the incorrectly assigned tags and missing named entities in the transcribed speech data. We assume that highly similar blocks of written data give extra knowledge about the named entities that are incorrectly as-

signed to the speech data and about the named entities missed in the speech data. The basic modeling work flow is composed of the following steps:

1. Transcribe the speech document using a common ASR system (FBK, 2013) and recognize the named entities in the speech document by a state-of-the-art NER tagger such as (Stanford, 2012). We will call the obtained list of unique named entities the *SNERList*.
2. Find related written texts. For instance, news sites could store related written texts with the broadcast video; or broadcast services might store speech and written data covering the same event. If that is not the case, written news data related to the given speech data might be crawled from the Web using some of the text similarity metrics or information retrieval systems. In the experiments below we choose the most related written document.
3. Divide the speech and written documents into fixed-size blocks. Each block contains n consecutive words. In the experiments below n = 50.[4]
4. Compute the similarity between the transcribed speech blocks and blocks of written text using the cosine similarity between their term vectors and align highly similar blocks. We call this step the *block alignment between speech and written data.*
5. If the similarity between a speech block and a block of written text is higher than a certain threshold, build a list of all named entities with their corresponding tags in the written text block again using the same NER tagger.
6. Group the unique named entities and their tags obtained from the aligned blocks of written text into the *WNERList*. This list contains valuable knowledge to update the *SNERList*
7. Correct and expand the *SNERList* based on the *WNERList*. The intuition is that we should trust the recognized named entities and their tags in the written data more than the ones obtained in the transcribed speech.

[4]We opt for aligning smaller chunks of information, that is, blocks instead of the entire documents. Incorrectly transcribed speech data introduce noise which negatively affects the quality of document alignment and, consequently, the overall NER system. The idea of working with only highly similar small blocks aims to circumvent the problem of noisy document alignments.

3.2 Our NER Models

The models that we propose differ in the ways they build the complete *SNERList* for a given speech document (Step 7 in the previous section) based on the knowledge in the *WNERList*.

3.2.1 Baseline NER Model

We use the Stanford NER on the transcribed speech data without any additional knowledge from similar written data. We call this model **Baseline NER**.

3.2.2 Correction and Expansion of the SNERList: General Principles

The procedure proceeds as follows: Let $(x_i)_{t_j}$ be the occurrence of the word x_i tagged by named entity class t_j in the *SNERList* and $(x_i)_{t_k}$ be the occurrence of the same word x_i now tagged by the named entity class t_k in the *WNERList*. Here, we assume the *one-sense-per-discourse-principle*, that is, all occurrences of the word x_i in a document can only belong to one NE class. We have to update the recognized named entities in the speech transcripts, i.e., replace $(x_i)_{t_j}$ with $(x_i)_{t_k}$ if it holds:

$$Count\big((x_i)_{t_j}\big) < Count\big((x_i)_{t_k}\big) \qquad (1)$$

The counts are computed in the related written document. This step is the *correction* of the *SNERList*. Additionally, we can expand the *SNERList* with named entities from the *WNERList* that were not present in the original *SNERList*. This step regards the *expansion* of the *SNERlist*.

3.2.3 Correction and Expansion of the SNERList Solely Based on the Edit Distance

The model updates the *SNERList* as follows. First, it scans the speech document and searches for orthographically similar words that are tagged in the similar written blocks. Orthographic similarity is modeled by the *edit distance* (Navarro, 2001). We assume that two words are similar if their edit distance is less than 2. The model not only uses the tags of the *WNERList* to correct the tags in the *SNERList* (see previous subsection), - we call this model **NER+COR**-, we also use newly linked words in the speech data to named entities of the *WNERList* to expand the *SNERList*. The model is called **NER+COR+EXP-ED**.

These models assign named entity tags only to words in the speech document that have their orthographically similar counterparts in the related written data. Therefore, they are unable to recover information that is missing in the transcribed speech document. Hence we need to design methods that expand the *SNERList* with relevant named entities from the written data that are missing in the transcribed speech document.

3.2.4 Expanding the SNERList with Named Entities from Written News Lead Paragraphs

It is often the case that the most prominent and important information occurs in the first few lines of written news (so-called *headlines* or *lead paragraphs*). Named entities occurring in these lead paragraphs are clearly candidates for the expansion of the *SNERList*. Therefore, we select named entities that occur in first 100 or 200 words in the related written news story and enrich the *SNERlist* with these named entities. Following that, we integrate the correction and expansion of named entity tags as before, i.e., this model is similar to **NER+COR+EXP-ED**, where the only difference lies in the fact that we now consider the additional expansion of the *SNERlist* by the named entities appearing in lead paragraphs. This model is called **NER+COR+EXP-ED-LP**.

3.2.5 Expanding the SNERList with Frequent Named Entities from Written News

The raw frequency of a named entity is a clear indicator of its importance in a written news document. Therefore, named entities occurring in related written documents are selected for expansion of the *SNERList* only if they occur at least M times in the written document on which the *WNERList* is based. Again, the correction part is integrated according to Eq. (1). We build the *SNERList* in the same manner as with the previous **NER+COR+EXP-ED-LP**, the only difference is that we now consider frequent words for the expansion of the *SNERlist*. This model is called **NER+COR+EXP-ED-FQ**.

3.2.6 Expanding the SNERList with Frequently Co-Occurring Named Entities from Written News

If a word in the related written document co-occurs many times with named entities detected in the original speech document, it is very likely that the word from the written document is highly descriptive for the speech document and should be taken into account for expansion of the *SNERlist*. We have designed three models that exploit the co-occurrence following an IR term expansion approach (Cao et al., 2008):

(i) Each word pair (s_i, w_j) consists of one named entity from the *SNERList* and one named entity from the *WNERList* that is currently not present in the *SNERList*. The co-occurrence is then modeled by the following formula:

$$SimScore_1(w_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\sum_B C(s_i, w_j|B)}{\sum_B tf(s_i, B)} \quad (2)$$

where $C(s_i, w_j|B)$ is the co-occurrence count of named entity s_i from the *SNERlist* and named entity w_j from the *WNERlist* not present in the former. The written document is divided into blocks and the co-occurrence counts are computed over all blocks B defined in section 3.1. $tf(s_i, B)$ is the frequency count of speech named entity s_i in block B. We call this model **NER+COR+EXP-ED-M1**.

(ii) The next model tracks the occurrence of each tuple (s_i, s_k, w_j) comprising two named entities from the *SNERlist* and one named entity w_j not present in the list, but which appears in the *WNERlist*. The co-occurrence is modeled as follows:

$$SimScore_2(w_j) = \sum_{(s_i,s_k)\epsilon\Omega} \sum_{j=1}^{n} \frac{\sum_B C(s_i, s_k, w_j|B)}{\sum_B tf(s_i, s_k, B)} \quad (3)$$

Again, $C(s_i, s_k, w_j|B)$ is the co-occurrence count of speech named entities s_i and s_j with named entity w_j in the written block B. Ω refers to all possible combinations of two named entities taken from the *SNERlist*. We call this model **NER+COR+EXP-ED-M2**.

(iii) The co-occurrence count in this model is weighted with the minimum distance between named entity s_i from the *S*NERList and named entity w_j that is a candidate for expansion. It assumes that words whose relative positions in the written document are close to each other are more related. Therefore, each pair is weighted conditioned on the distance between the words in a pair. The distance is defined as the number of words between two words. The co-occurrence score is then computed as follows:

$$SimScore_3(w_j) = \frac{\sum_B \frac{C(s_i,w_j)}{minDist(s_i,w_j)}}{\sum_B C(s_i,w_j)} \quad (4)$$

where $minDist(s_i, w_j)$ denotes the minimum distance between words s_i and w_j. The model is **NER+COR+EXP-ED-M3**.

These 3 models are similar to the other models that perform the expansion of the *SNERlist*. The difference is that the expansion is performed only with candidates from the *WNERlist* that frequently co-occur with other named entities from the *SNERlist*.

4 Experimental Setup

4.1 Datasets and Ground Truth

For evaluation we have downloaded 11 short broadcast news from the Internet (the sources are `tv.msnbc.com` and `www.dailymail.co.uk`). The FBK ASR transcription system (FBK, 2013) is used to provide the speech transcriptions from the data. Since the system takes sound as input, we have extracted the audio files in the mp3 format using the ffmpeg tool (ffm, 2012). The transcribed speech data constitute our *speech dataset*. The following table shows an example of a manual transcription and the transcription outputed by the FBK ASR system. The speech documents need to be labeled with 143 unique named entities and their named entity tag.

Manual Transcription
hbo is debuting a documentary on thursday night about the very public american life of ethel kennedy. at 84, she's certainly a survivor having lost her husband and two of her 11 children. she has seldom spoken about her life about any of it. she's talked to us about the parts of her story she's chosen to reveal now. i mean, how lucky could i really have been.
ethel kennedy makes a big distinction between counting her blessings and not looking back too much. while she does not enjoy being interviewed, she said yes to her daughter rory, a documentary filmmaker who wanted to tell her mom's extraordinary story.....

ASR Transcription
HBO is debuting a documentary on Thursday night about the very public American wife of Ethel Kennedy any for she is certainly a survivor having lost her husband and two or for another job she has seldom spoken of our life about any other when she talked to us about the Passover story she's chosen to reveal now in highlighting Ethel Kennedy makes a big distinction between counting their blessings and not looking back to March washing does not enjoy being interviewed she said yesterday over borrowing a documentary film maker who wanted to calls.....

Figure 1: An example of the actual transcription done manually and the transcription done by the FBK ASR system.

Fig. 1 shows that the ASR transcription contains many words that are incorrectly transcribed. It is also visible that the ASR system does not recognize and misspells many words from the actual speech.

The related written news stories of the 11 broadcast news are collected from different news sources available on the Web such as `http://www.guardian.co.uk`, `http://www.independent.co.uk`, `www.cnn.com`, etc. The collected written news stories constitute our *written text dataset*.

In order to build the ground truth for our experiments, all 11 stories were manually transcribed. Stanford NER was then applied on the manually transcribed data. Following that, the annotator checked and revised the NER-tagged lists. The ground truth was finally created by retaining the revised lists of named entities with their corresponding tags. We work with the following 3 common named entity tags: *person*, *location* and *organization*.

4.2 Evaluation Metrics

Let FL be the final list of named entities with their corresponding tags retrieved by our system for all speech documents, and GL the complete ground truth list. We use standard precision ($Prec$), recall (Rec) and F-1 scores for evaluation:

$$Prec = \frac{|FL \cap GL|}{|FL|} \qquad Rec = \frac{|FL \cap GL|}{|GL|}$$

$$F_1 = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

We perform *evaluation at the document level*, that is, we disregard multiple occurrences of the same named entity in it. In cases when the same named entity is assigned different tags in the same document (e.g., *Kerry* could be tagged as *person* and as *organization* in the same document), we penalize the system by always treating it as an incorrect entry in the final list FL.

This evaluation is useful when one wants to index a speech document as a whole and considers the recognized named entities and their tags as document metadata. Within this evaluation setting it is also possible to observe the models' ability to recover missed named entities in speech data.

4.3 Parameters

The notion of "frequent co-occurrence" is specified by a threshold parameter. Only words that score above the threshold are used for expansion. Based on a small validation set of two speech documents and their corresponding written document, we set the threshold value for **NER+COR+EXP-ED-M1** and **NER+COR+EXP-ED-M2** to 0.01, while it is 0.002 for **NER+COR+EXP-ED-M3**. All results reported in the next section are obtained using these parameter settings, but by fluctuating

NER Model	Precision	Recall	F-1
Baseline NER	**0.407**	**0.567**	**0.474**
NER+COR	0.427	0.594	0.497
NER+COR+EXP-ED	0.411	0.601	0.489
NER+COR+EXP-ED-LP ($\|LP\| = 100$)	0.359	0.678	0.470
NER+COR+EXP-ED-LP ($\|LP\| = 200$)	0.322	0.678	0.437
NER+COR+EXP-ED-FQ ($M = 2$)	0.387	0.657	0.487
NER+COR+EXP-ED-FQ ($M = 3$)	0.411	0.650	0.504
NER+COR+EXP-ED-M1	**0.415**	**0.650**	**0.507**
NER+COR+EXP-ED-M2	0.414	0.622	0.497
NER+COR+EXP-ED-M3	0.384	0.664	0.487

Table 1: Results of different NE recovering models on the evaluation dataset.

them precision increases while recall decreases, or vice versa.

5 Results and Discussion

Table 1 shows all the results of our experiments, where we compare our models to the baseline model that uses the named entity recognizer for tagging the speech data, i.e., Baseline NER. We may observe that our system is able to correct the tag of some named entities in the transcribed speech data by the **NER+COR** model and expand some missed named entities by the **NER+COR+EXP-ED** model. All models are able to recover a subset of missing named entities, and that fact is reflected in increased recall scores for all models. The **NER+COR+EXP-ED-M1** model outperforms the other models and improves the F1 by 3% with an increase in 8% in recall and almost 1% in precision.

In our dataset there are 27 unique named entities that are in the ground truth transcription of the speech data, but are missing completely in the transcribed speech data. Out of these 27 named entities, 8 named entities do not occur in the written related documents, so we cannot learn these from the written data. Out of 19 named entities recoverable from written data our system is able to correctly identify 6 named entities and their tags with the **NER+COR+EXP-ED-M1** model. We can lower the threshold for the similarity score computed in Eq. (3). For instance, when we substantially lower the threshold to 0.001 we correctly find 12 missing named entities, but the increased recall is at the expense of a much lower precision ($P = 0.290, R = 0.699, F1 = 0.411$), because many irrelevant named entities are added to the final *SNERList*. We have also investigated why the remaining 7 named entities seen in the written data are not recovered even with such a low threshold. We noticed that those named entities do not co-occur with the named entities found in the speech transcripts in the considered blocks of the written texts. Hence, our methods can still be improved by finding better correlations between named entities found in the speech and related written documents. The named entity recognition in the related written texts is not perfect either and can entail errors in the corrections and expansions of the named entities found in the speech data.

6 Conclusions and Future Work

In this paper we have shown that NER from speech data benefits from aligning broadcast news data with related written news data. We can both correct the identified named entity tags found in the speech data and expand the named entities and their tags based on knowledge of named entities from related written news. The best improvements in terms of precision and recall of the NER are obtained with word expansion techniques used in information retrieval. As future work we will refine the named entity expansion techniques so to further improve recall and to better capture missing named entities without sacrificing precision, we will consider several speech transcription hypotheses, and we will try to improve the named entity recognition itself by making the models better portable to texts that are different from the ones they are trained on.

7 Acknowledgements

This research was supported by the EXPERTS Erasmus Mundus Action 2 scholarship of the first author and by the EU FP7 project TOSCA-MP No. 287532.

References

Roberto Basili, Marco Cammisa, and Emanuale Donati. 2005. RitroveRAI: A Web application for semantic indexing and hyperlinking of multimedia news. In *Proceedings of International Semantic Web Conference*, pages 97–111.

Frédéric Béchet, Allen L Gorin, Jeremy H Wright, and Dilek Hakkani Tur. 2004. Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may i help you? *Speech Communication*, 42(2):207 – 225.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–250.

Nancy A. Chinchor. 1997. MUC-7 named entity task definition (version 3.5). In *Proceedings of the 7th Message Understanding Conference*.

Benoît Favre, Frédéric Béchet, and Pascal Nocera. 2005. Robust named entity extraction from large spoken archives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 491–498.

FBK. 2013. FBK ASR transcription.

2012. ffmpeg audio/video tool @ONLINE.

James Horlock and Simon King. 2003. Discriminative methods for improving named entity extraction on speech data. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 2765–2768.

Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 1998. Named entity extraction from speech. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 287–292.

David Miller, Richard Schwartz, Ralph Weischedel, and Rebecca Stone. 1999. Named entity extraction from broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*, pages 37–40.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

David D. Palmer and Mari Ostendorf. 2001. Improving information extraction by modeling errors in speech recognizer output. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–5.

Mark A. Przybocki, Jonathan G. Fiscus, John S. Garofolo, and David S. Pallett. 1999. HUB-4 information extraction evaluation. In *Proceedings of the DARPA Broadcast News Workshop*, pages 13–18.

Lisa F. Rau. 1991. Extracting company names from text. In *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications*, pages 29–32.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-Independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 1–4.

Stanford. 2003. Stanford NER in CoNLL 2003.

Stanford. 2012. Stanford NER.

Beth Sundheim. 1995a. Named entity task definition. In *Proceedings of the 6th Message Understanding Conference*, pages 317–332.

Beth Sundheim. 1995b. Overview of results of the MUC-6 evaluation. In *Proceedings of the 6th Message Understanding Conference*, pages 13–31.