An IR-Inspired Approach to Recovering Named Entity Tags in Broadcast News

Niraj Shrestha, Ivan Vulić and Marie-Francine Moens

Department of Computer Science, KU Leuven, Belgium
{niraj.shrestha,ivan.vulic,marie-francine.moens}@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 automatically detect document alignments between highly similar speech documents and corresponding written news stories 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 named entities missing in the transcribed speech data, and additionally to correct incorrectly assigned named entity tags. Consequently, our novel approach improves state-of-the-art NER results from speech data both in terms of recall and precision.

Keywords: Named entity recognition, term expansion, broadcast news, speech data

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., [17]). The task involves labeling (proper) nouns with suitable *named entity tags*, and it is usually treated as a sequence prediction problem. NER is an important pre-processing task in many applications in the fields of information retrieval (IR) and natural language processing.

NER in speech data also displays its utility in various multimedia applications. For instance, it could be used in indexing video broadcast news using associated speech data, that is, assigning names and their semantic classes recognized from the speech data as metadata to the video sequences [2]. It is also a useful component of speech-based question answering systems (e.g., [16]), or it could be used to extract names from meeting minutes provided in audio format.

NER in speech data is a difficult task and current state-of-the-art results are typically much lower than the results obtained in 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% [23]. [13, 15] 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 reasons. Firstly, speech transcribers often incorrectly transcribe phrases and even complete sentences, which might consequently result in many missing named entities. Secondly, many names are typically not observed in the training data on which the speech transcriber is 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 spelled incorrectly, making the named entity recognition even more challenging. Finally, the NER system, 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 documents easily obtainable from the Web 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 are selected as being relevant for the query by the user or automatically by the IR system [6]. Only a subset of terms is selected for expansion based on their importance in the relevant document, as well as their 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 related relevant written documents. Following this modeling intuition, we are able to improve the recall of the NER from broadcast speech data by almost 9%, while precision scores increase for around 0.4% compared to the results of applying the same named entity recognizer on the speech data directly. The main contributions of this article are as follows:

1. We show that NER from speech data benefits from aligning broadcast news data with similar written news data.
2. We present several new methods to recover named entities from speech data by using the external knowledge from high-quality similar written texts.
3. We improve the performance of the state-of-the-art Stanford NER system when applied to the transcribed speech data. The utility of the recovering of missing named entities is especially prominent in much higher recall scores, while we manage to retain a stable and even slightly improved precision level.

The following sections first review prior work, then describe the methodology of our approach and the experimental setup, and finally present our evaluation procedure and discuss the results.

2 Prior Work

There exists a significant body of work on named entity recognition in written data. The task was initially defined in the framework of the Message Understanding Conferences (MUC) [24]. Since then, many conferences and workshops such as the following MUC editions [24, 7], the 1999 DARPA broadcast news workshop [21] and the CoNLL shared tasks [22] focused on extending state-of-the-art research on NER.

The most common approach to named entity recognition is based on word-by-word sequential classification techniques, similar to the techniques frequently used for part-of-speech tagging and syntactic base-phrase chunking. A classifier is trained to label each word token in an input text one after the other, in sequence, using the appropriate named entity tag. Current state-of-the-art NER models typically rely on machine learning algorithms and probabilistic hidden state sequence models such as Hidden Markov

Models, Maximum Entropy Markov Models or Conditional Random Fields trained on documents with manually annotated named entities. A myriad of NER implementations are widely available. Examples include the Stanford NER system[1], OpenNLP NameFinder[2], Illinois NER system[3], and LingPipe NER[4]. In this work we utilize the Stanford Named Entity Recognizer, because of its state-of-the-art results, accessible source code and user-friendly interface.

The Stanford NER system [10] identifies named entities of four different types, *person*, *location*, *organization*, and *miscellaneous*.[5] The system recognizes named entities using a combination of three linear chain Conditional Random Field (CRF) sequence taggers. The features used are, among others, word features based on the words in the context window, such as the words themselves and their part-of-speech, orthographic features, prefixes and suffixes of the word to be labeled and surrounding words, and distributional similarity based features. The CRF sequence models are trained on a mixture of various corpora with manually annotated named entities, such as CoNLL, MUC-6 and MUC-7 corpora. These corpora contain both British and American newswire articles, so the resulting models should be fairly robust across domains.

Unfortunately, when applying such a state-of-the-art NER system on transcribed speech data, the performance deteriorates dramatically. 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 [24, 13].

NER in speech data was initiated by [13]. He applied a NER system on transcriptions of broadcast news, and reported that its performance degraded linearly with the word error rate of speech recognition (e.g., missing data, misspelled data and spuriously tagged names). Named entity recognition in speech data has been investigated further, but this related work has focused on either decreasing the error rate when transcribing speech [15, 20], on considering different speech transcription hypotheses [11, 3], or on the issue of temporal mismatch between training and test data [8]. None of these articles consider exploiting external text sources to improve NER in speech data nor the problem of recovering missing named entities in transcribed speech. Another line of work [5, 14] has proven that performing a lexical expansion using related written text obtainable from the Web may boost the performance of systems for speech language modeling, but none of the prior work performed the expansion from written Web sources in the task of NER in speech data. [4, 19] link video news stories with written news data. They used closed captions or sub-titles to search related written stories, but do not report on recovering missing named entities.

[1] `http://nlp.stanford.edu/software/stanford-ner-2012-11-11.zip`

[2] `http://opennlp.sourceforge.net/models-1.5`

[3] `http://cogcomp.cs.illinois.edu/page/software_view/4`

[4] `http://alias-i.com/lingpipe/web/models.html`

[5] The system is also able to recognize numerical entities of types *date*, *time*, *money*, and *number*, but we are interested only in the first 3 basic types.

3 Recovering Named Entity Tags in Speech: Methodology

The task is to label a sequence of words $[w_1, w_2, \ldots, w_N]$ from transcribed broadcast news data 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 case of the Stanford NER system utilized in this work, $t_i \in \{person, organization, location\}$.

3.1 Basic Architecture

The straightforward approach to NER in speech data in prior work is to apply a state-of-the-art text data NER tagger (e.g., Stanford NER) directly on transcribed speech data. However, the tagger will miss many named entities or assign incorrect named entity (NE) tags due to the inherent errors in the speech 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 written documents or blocks of texts give extra knowledge about the named entities that are incorrectly assigned to the speech data and about the named entities missed in the speech data. The basic modeling work flow follows these steps:

1. Transcribe the speech document using a state-of-the-art ASR system [9] and recognize the named entities in the speech document by a state-of-the-art NER tagger. We call the list of unique named entities obtained in this initial step the *SNERList*.
2. Find related written texts. For instance, news sites often store related written texts with the broadcast video (e.g., Google news). Written news related to the given speech data might also be automatically crawled from the Web. In both cases we use a text similarity metric (e.g., the cosine similarity) to identify related written texts.
3. Group the unique named entities and their tags obtained from the related documents or aligned blocks of written text into the *WNERList*. This list contains valuable knowledge that is utilized to update the *SNERList*.
4. Correct and expand the *SNERList* based on the *WNERList* forming a final list of named entities called FL, the named entities of which can be used as metadata for indexing the speech document. The intuition here is that we should trust the recognized named entities and their tags in the written data more than in the corresponding transcribed speech data.

The models that we propose below differ in the manner they build the complete *SNERList* for a given speech document (Step 4) based on the knowledge in the *WNERList*.

3.2 Baseline NER Model

As a baseline model, we use the Stanford NER system applied on transcribed speech data without any additional knowledge coming from similar written data. We call this model **Baseline NER**.

3.3 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 NE class t_j in the *SNERList* and $(x_i)_{t_k}$ be the occurrence of the same word x_i

now tagged by the NE 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 may 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:

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

The counts are computed in the most related written document computed in step 2 of the above procedure. This step regards the *correction* of the *SNERList*. This first model that uses the tags of the *WNERList* to correct the *SNERList* is called **NER+COR**. Additionally, we can expand the *SNERList* with named entities from the *WNERList* that were not present in the original *SNERList*. This step denotes the *expansion* of the *SNERList*, but we need to design a smart strategy of selecting named entities from written text that are suitable for the expansion.

3.4 Correction and Expansion of the SNERList 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 of the most related written document computed in steps 2 and 3 of the above procedure. Orthographic similarity is modeled by the *edit distance* [18]. We assume that two words are similar if their edit distance is less than 2. The model is similar to NER+COR, but it additionally utilizes orthographic similarity to link words in the speech data to named entities in the *WNERList* in order to expand the *SNERList*. The model is called **NER+COR+EXP-ED**.

These models assign NE 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 additional methods that further expand the *SNERList* with relevant named entities from the written data that are missing in the transcribed speech document. Below we list several alternative approaches to accomplish this goal.

3.5 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 the first 100 or 200 words in the most related written news story and enrich the *SNERList* with these named entities. Following that, we integrate the correction and expansion of NE 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**.

3.6 Expanding the SNERList with Frequent Named Entities from Written News

The raw frequency of a NE is also a straightforward indicator of its importance in a written news document. Therefore, named entities are selected for expansion of the

SNERList if they occur at least M times in the most related written document used to build the *WNERList*. 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 model, 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**.

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

If a NE in the most related written document co-occurs many times with NEs detected in the original speech document, it is very likely that this NE 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 [6].
We compute a score ($SimScore$) for each NE (w_j) in the *WNERList* by which w_j can be ranked according to its relevance for the speech document represented as the set of NEs of the *SNERList*. This co-occurrence score is then modeled in three variant models. The first two models consider the co-occurrence of the NE s_i in the speech document and w_j in blocks of n consecutive words in the written document. So the written document is divided in x blocks B_l. In the third model the distance between s_i and w_j in the written document is taken into account.

(i) Each entity pair (s_i, w_j) consists of one NE from the *SNERList* and one NE from the *WNERList* that is currently not present in the *SNERList* and which is thus a candidate for expansion.

$$SimScore_1(w_j) = \frac{1}{v} \sum_{s_i \in SNERList} \frac{\sum_{B_l} C(s_i, w_j | B_l)}{\sum_{w_k \in WNERList} \sum_{B_l} tf(w_k, B_l)} \qquad (2)$$

where $C(s_i, w_j|B_l)$ is the co-occurrence count of NE s_i from the *SNERList* and NE w_j in the written text. The co-occurrence counts are computed over all blocks. $tf(w_k, B_l)$ is the frequency count of the NE w_k in block B_l. We call this model **NER+COR+EXP-ED-M1**. We average the scores over all s_i of the *SNERList*, where v is the number of NEs in the *SNERList*. In a variant model we have normalized the co-occurrence counts of s_i and w_j in block B_l with the co-occurrence counts of s_i with any w_k in block B_l resulting in a very similar performance.
(ii) The next model tracks the occurrence of each tuple (s_i, s_z, w_j) comprising two named entities from the *SNERList* and one NE w_j not present in the list, but which appears in the *WNERList*. The co-occurrence is modeled as follows:

$$SimScore_2(w_j) = \frac{1}{|\Omega|} \sum_{(s_i, s_z) \epsilon \Omega} \frac{\sum_{B_l} C(s_i, s_z, w_j | B_l)}{\sum_{w_k \in WNERList} \sum_{B_l} tf(w_k, B_l)} \qquad (3)$$

Again, $C(s_i, s_z, w_j|B)$ is the co-occurrence count of speech named entities s_i and s_z with NE w_j in the written block B_l. Ω refers to all possible combinations of two NEs 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 NE s_i from the *S*NERList and NE 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 entities in a pair. The distance is defined as the number of words between the two NEs. The co-occurrence score is then:

$$SimScore_3(w_j) = \frac{\sum_{s_i \in SNERList} \sum_{B_l} \frac{C(s_i, w_j | B_l)}{minDist(s_i, w_j)}}{\sum_{s_i \in SNERList} \sum_{B_l} C(s_i, w_j | B_l)} \tag{4}$$

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

These 3 models are similar to the other models that perform the expansion of the *S*NERList. The difference is that the expansion is performed only with candidates from the *WNERList* that frequently co-occur with other named entities from the *S*NERList. The notion of "frequent co-occurrence" is specified by a threshold parameter and only entities that score above the threshold are used for expansion.

3.8 Expanding the SNERList with Intersection between Named Entities from a Set of Related Written News Documents

The idea of this model is to retain only the named entities that occur in many related news documents (computed in step 2 of the above procedure) as candidates for the expansion of the *SNERList*. Here, we first select a set of related written news text for each speech document, and then, to expand the *SNERList*, we use the intersection of the named entities, i.e., named entities that occur in all written documents in the corresponding set. We can select the related written documents based on a minimum similarity value, or based on a cut-off in the ranked list of related written documents. Selecting a minimum similarity value is not straightforward. If we take a very low similarity score, it might introduce one or more unrelated written documents. In that case, we might end up with an empty intersection list. If we choose a high similarity score, we might lose relevant related written documents. This model is named as **NER+COR+EXP-ED-INTS**. The selection of K related documents is easier. We might even choose the K related written stories that can be found on the same event on a given date, where K is a flexible number. In the experiments below we use a fixed number K for all speech examples. This model is called as **NER+COR+EXP-ED-INTS-BK**.

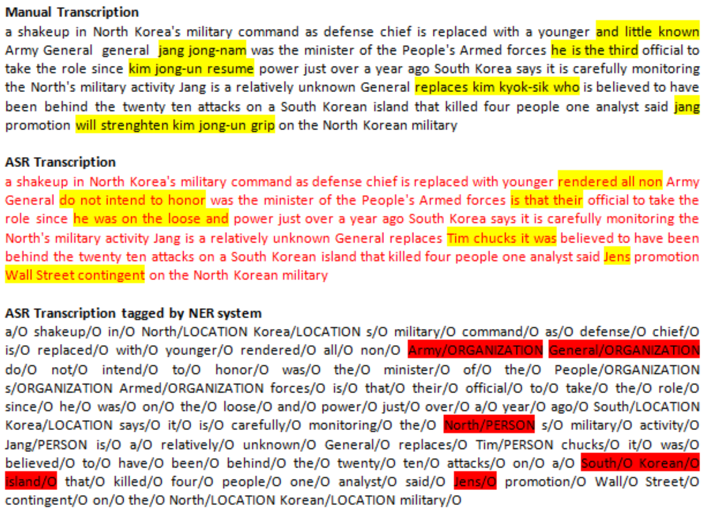
All the above models, some of which are borrowed from query expansion research in IR, show the many possible ways of exploiting the named entities in written documents that are related to the speech document in which we want to improve NER.

4 Experimental Setup

4.1 Datasets and Ground Truth

For evaluation, we have downloaded 40 short broadcast news stories from the Web in the periods of October-November 2012 and April-May 2013 randomly selected from `www.googlenews.com`, `tv.msnbc.com`, `bbc.com`, `cnn.com`, and `www.dailymail.co.uk`.[6] We have collected 5532 related news stories from `www.news.google.com` which stores related news stories from different sites, and they constitute our

[6] Dataset is availabel at `http://people.cs.kuleuven.be/˜niraj.shrestha/NER`



Manual Transcription

a shakeup in North Korea's military command as defense chief is replaced with a younger and little known Army General general jang jong-nam was the minister of the People's Armed forces he is the third official to take the role since kim jong-un resume power just over a year ago South Korea says it is carefully monitoring the North's military activity Jang is a relatively unknown General replaces kim kyok-sik who is believed to have been behind the twenty ten attacks on a South Korean island that killed four people one analyst said jang promotion will strenghten kim jong-un grip on the North Korean military

ASR Transcription

a shakeup in North Korea's military command as defense chief is replaced with younger rendered all non Army General do not intend to honor was the minister of the People's Armed forces is that their official to take the role since he was on the loose and power just over a year ago South Korea says it is carefully monitoring the North's military activity Jang is a relatively unknown General replaces Tim chucks it was believed to have been behind the twenty ten attacks on a South Korean island that killed four people one analyst said Jens promotion Wall Street contingent on the North Korean military

ASR Transcription tagged by NER system

a/O shakeup/O in/O North/LOCATION Korea/LOCATION s/O military/O command/O as/O defense/O chief/O is/O replaced/O with/O younger/O rendered/O all/O non/O Army/ORGANIZATION General/ORGANIZATION do/O not/O intend/O to/O honor/O was/O the/O minister/O of/O the/O People/ORGANIZATION s/ORGANIZATION Armed/ORGANIZATION forces/O is/O that/O their/O official/O to/O take/O the/O role/O since/O he/O was/O on/O the/O loose/O and/O power/O just/O over/O a/O year/O ago/O South/LOCATION Korea/LOCATION says/O it/O is/O carefully/O monitoring/O the/O North/PERSON s/O military/O activity/O Jang/PERSON is/O a/O relatively/O unknown/O General/O replaces/O Tim/PERSON chucks/O it/O was/O believed/O to/O have/O been/O behind/O the/O twenty/O ten/O attacks/O on/O a/O South/O Korean/O island/O that/O killed/O four/O people/O one/O analyst/O said/O Jens/O promotion/O Wall/O Street/O contingent/O on/O the/O North/LOCATION Korean/LOCATION military/O

Fig. 1. An example of the actual transcription performed manually, the transcription obtained by the FBK ASR system and the ASR transcription tagged by the Stanford NER system.

Table 1. Statistics of 40 broadcast news data used for evaluation.

	Frequency of named entities
# **NEs in ground truth**	408
# **NEs transcribed** by FBK ASR	302
# **NEs not transcribed** by FBK ASR (missing names)	106
# **NEs tagged** by Stanford NER	487
# **NEs correctly tagged** by Stanford NER	283
# **NEs incorrectly tagged** by Stanford NER	204

written text dataset. The FBK ASR transcription system [9] is used to provide the speech transcriptions of these stories. Since the system takes sound as input, we have extracted the audio files in the mp3 format using the *ffmpeg* tool [1]. The transcribed speech data constitute our *speech dataset*. Fig. 1 shows an example of the manual transcription, and its tagging by the Stanford NER system.It is clear that the ASR transcription contains many words that are incorrectly transcribed and that the ASR system does not recognize many words from the actual speech. It is also noted that the NER system could not tag the missed named entities in the ASR transcription.

In order to build the ground truth for our experiments, all 40 broadcast news stories were manually transcribed. The Stanford NER was then applied on the manually transcribed data. Following that, an annotator checked and revised the tagged named entities. The detailed statistics are provided in Table 1. There are all together 106 named entities missing from the speech data set due to transcription errors, and these cannot be tagged by the NER system. Additionally, we observe that a large portion of the named entities is incorrectly tagged by Stanford NER. The knowledge from aligned written data should help us resolve these issues.

Table 2. Results of different NE recovering models on the evaluation dataset.

NER Model	Precision	Recall	F_1
Baseline NER	**0.508**	**0.605**	**0.553**
NER+COR	0.521	0.620	0.566
NER+COR+EXP-ED	0.506	0.632	0.562
NER+COR+EXP-ED-LP ($\|LP\| = 100$)	0.444	0.706	0.545
NER+COR+EXP-ED-LP ($\|LP\| = 200$)	0.393	0.718	0.508
NER+COR+EXP-ED-FQ ($M = 2$)	0.438	0.686	0.535
NER+COR+EXP-ED-FQ ($M = 3$)	0.490	0.674	0.568
NER+COR+EXP-ED-M1	0.518	0.634	0.570
NER+COR+EXP-ED-M2	0.516	0.632	0.568
NER+COR+EXP-ED-M3	0.377	0.662	0.480
NER+COR+EXP-ED-INTS-BK ($K = 3$)	0.479	0.725	0.577
NER+COR+EXP-ED-INTS-BK ($K = 5$)	**0.512**	**0.694**	**0.589**
NER+COR+EXP-ED-INTS-BK ($K = 7$)	0.517	0.662	0.581
NER+COR+EXP-ED-INTS-BK ($K = 10$)	0.520	0.642	0.575

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:

$$Prec = \frac{|FL \cap GL|}{|FL|} \quad Rec = \frac{|FL \cap GL|}{|GL|} \quad F_1 = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

We perform an *evaluation at the document level*, that is, we disregard multiple occurrences of the same named entity in one document. 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.

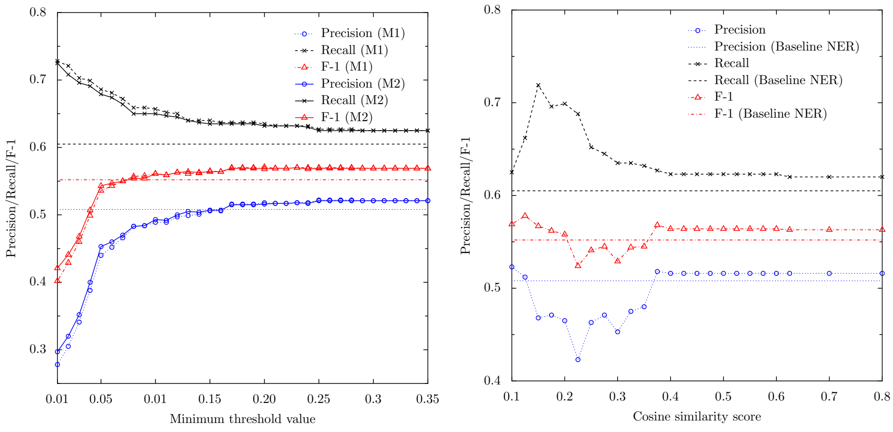
5 Results and Discussion

Table 2 displays all results obtained on the evaluation dataset of 40 broadcast news stories. We compare the results of our models to the baseline model that uses a NER system directly on transcribed speech data (Baseline NER). We observe that all the proposed models are able to correct a portion of the named entities initially missed in the transcribed speech data (for instance, we notice a performance small boost already by using the simple NER+COR model), and additionally expand the *SNERList* by named entities from similar written data (see performance boosts, especially boosts in terms of recall for all models that perform the *SNERList* expansion). A majority of the proposed models outperform the baseline NER system in terms of F_1 score, and they all exhibit significant performance boosts in terms of recall. The best results are obtained by the NER+COR+EXP-ED-INTS-BK model with $K = 5$, where we can observe an increase of 9% in terms of recall (due to the expansion procedure) (significant for $p < 0.0002$ 2-paired t-test), and a 0.4% increase in terms of precision (significant for $p < 0.14$) that is altogether reflected in a 3.6% increase in terms of F_1 score.

We have investigated the influence of the minimum threshold values on the results obtained by the term co-occurrence models (NER+COR+EXP-ED-M1/M2). Figure 2(a) displays the dependence on the threshold value. We observe that by setting a low threshold we are able to recover a considerable number of named entities (a 12% increase in recall for the threshold of 0.01), but it degrades the precision scores. The best results presented in Table 2 are obtained by the threshold value of 0.2.

We have also investigated the influence of the minimum cosine similarity score that is needed to consider a written document similar to the given speech document in the NER+COR+EXP+ED+INTS model. The results are displayed in Fig. 2(b). If we choose a lower similarity threshold then the model tends to select unrelated written documents as similar to the given speech document and it has a negative impact on the overall results. On the other hand, if the selected threshold is set too high, the model tends to omit relevant related written documents. Results in Table 2 are obtained by setting the similarity threshold to 0.125, but we observe a stable performance for other threshold settings. Similarly, for the K-best cut-off intersection model, we have varied the cut-off position ($K = 3, 5, 7, 10$). The results are displayed in Table 2. Since the F_1 score is stable for different K values, it confirms our hypothesis that K might be chosen in a flexible way, for instance, as the number of written documents on the same event reported in the speech document available on a certain day.

To recover the missing named entities, the system should learn from the related written text. Out of 106 missing named entities (see the statistics in Table 1), there are only 89 named entities observed in the related written news dataset. This constitutes



(a) (b)

Fig. 2. Influence of parameters on the overall results: (a) threshold value for the term co-occurrence models NER+COR+EXP-ED-M1 and NER+COR+EXP-ED-M2, (b) minimum similarity value for the intersection model NER+COR+EXP-ED-INTS.

the upper bound of our approach. In order to deal with this problem, we need to collect more related news stories. Furthermore, we have noticed that out of 17 missing named entities, NEs like news anchor or reporter names can rarely be found in related written text. Our best model NER+COR+EXP-ED-INTS-BK with $K = 5$ recovers 31 named entities out of 106 missing named entities from the related news texts boosting the recall substantially without hurting precision.

We are able to recover a substantially larger amount of missing named entities by lowering the threshold for the similarity score computed in Eq. (2) and (3) in models NER+COR+EXP-ED-M1/M2. For instance, as shown in the Fig. 2(a) when we lower the threshold to 0.01, the recall increases to 72.8% and the system recovers 53 missing named entities, but the increased recall is at the expense of a much lower precision ($P = 27.78\%$). In that setting many irrelevant named entities are added to the *SNERList*. Our methods can still be improved by finding better correlations between named entities found in the speech and related written documents. One line of our future research will strive to retain the substantial increases in terms of recall while retaining a stable precision level.

The NE recognition in the related written texts is not perfect either and can entail errors in the correction and expansion of the named entities found in the speech data. [12] report that the performance of the Stanford NER system in Web data decreases by 14%. To confirm this finding, we have also checked the performance of Stanford NER when applied on our written text. We have randomly selected 20 written news stories and run the Stanford NER. The performance is ($P = 76.69\%, R = 80.89\%, F_1 = 78.73\%$) which clearly indicates that there is still ample room for improvement in the task of NER in written data. Further improvements in NER in written data will also have a positive impact on the models for NER in speech data that we propose in this article.

6 Conclusions and Future Work

We have proposed a novel IR-inspired approach to recovering NE tags in transcribed speech using similar written texts. We have shown that NER from speech data benefits from aligning broadcast news data with related written news data. Our new models are able to both (1) correct tags for named entities identified in the speech data that were tagged incorrectly, and (2) expand the list of named entities in the speech data based on the knowledge of named entities from related written news stories. The best improvements in terms of precision and recall of the NER are obtained by considering the named entities that occur in the intersection of several related written documents. Our results show that we can improve the recall of the NER by 9% compared to solely considering NER in the transcribed speech data without hurting precision. In our evaluation dataset almost 25% of named entities were missing after the ASR transcription, and we have shown that our best method is able to correctly recover and tag almost one third of the missing named entities.

In future work we plan to further refine the NE expansion techniques in order to enrich the lists of named entities in speech using written data without sacrificing precision. We also plan to explore several other speech transcription hypotheses, and study the core problem of domain adaptation when dealing with the task of NE recognition in order to build more portable NER taggers.

References

1. ffmpeg audio/video tool @ONLINE (2012), `http://www.ffmpeg.org`
2. Basili, R., Cammisa, M., Donati, E.: RitroveRAI: A Web application for semantic indexing and hyperlinking of multimedia news. In: Proc. of ISWC. pp. 97–111 (2005)
3. Béchet, F., Gorin, A.L., Wright, J.H., Tur, D.H.: Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may I help you? Speech Comm. 42(2), 207–225 (2004)
4. Blanco, R., De Francisci, Morales, G., Silvestri, F.: Towards leveraging closed captions for news retrieval. In: Proc. of WWW companion. pp. 135–136 (2013)
5. Bulyko, I., Ostendorf, M., Stolcke, A.: Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In: Proc. of NAACL-HLT. pp. 7–9 (2003)
6. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proc. of SIGIR. pp. 243–250 (2008)
7. Chinchor, N.A.: MUC-7 named entity task definition (version 3.5). In: Proc. of MUC (1997)
8. Favre, B., Béchet, F., Nocera, P.: Robust named entity extraction from large spoken archives. In: Proc. of EMNLP. pp. 491–498 (2005)
9. FBK: FBK ASR transcription (2013), `https://hlt-tools.fbk.eu/tosca/publish/ASR/transcribe`
10. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proc. of ACL. pp. 363–370 (2005)
11. Horlock, J., King, S.: Discriminative methods for improving named entity extraction on speech data. In: Proc. of EUROSPEECH. pp. 2765–2768 (2003)
12. Kim, M.H., Compton, P.: Improving the performance of a named entity recognition system with knowledge acquisition. In: Proc. of EKAW. pp. 97–113 (2012)
13. Kubala, F., Schwartz, R., Stone, R., Weischedel, R.: Named entity extraction from speech. In: Proc. of the DARPA Broadcast News Transcription and Understanding. pp. 287–292 (1998)
14. Lei, X., Wang, W., Stolcke, A.: Data-driven lexicon expansion for Mandarin broadcast news and conversation speech recognition. In: Proc. of ICASSP. pp. 4329–4332 (2009)
15. Miller, D., Schwartz, R., Weischedel, R., Stone, R.: Named entity extraction from broadcast news. In: Proc. of the DARPA Broadcast News. pp. 37–40 (1999)
16. Mishra, T., Bangalore, S.: Qme! : A speech-based question-answering system on mobile devices. In: Proc. of NAACL-HLT. pp. 55–63 (2010)
17. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
18. Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys 33(1), 31–88 (2001)
19. Odijk, D., Meij, E., de Rijke, M.: Feeding the second screen: semantic linking based on subtitles. In: Proc. of the 10th Conference on OAIR. pp. 9–16. OAIR '13 (2013)
20. Palmer, D.D., Ostendorf, M., Burger, J.D.: Robust information extraction from automatically generated speech transcriptions. Speech Comm. 32(1-2), 95–109 (2000)
21. Przybocki, J.M., Fiscus, J.G., Garofolo, J.S., Pallett, D.S.: HUB-4 information extraction evaluation. In: Proc. of the DARPA Broadcast News. pp. 13–18 (1999)
22. Sang, E.F.T.K., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: Language-Independent named entity recognition. In: Proc. of CoNLL. pp. 142–147 (2003)
23. Stanford: Stanford NER in CoNLL 2003 (2003), `http://nlp.stanford.edu/projects/project-ner.shtml`
24. Sundheim, B.: Overview of results of the MUC-6 evaluation. In: Proc. of MUC. pp. 13–31 (1995)