Entity Linking for Spoken Language

Adrian Benton
Human Language Technology
Center of Excellence
Johns Hopkins University
Baltimore, MD, USA
adrian@jhu.edu

Mark Dredze
Human Language Technology
Center of Excellence
Johns Hopkins University
Baltimore, MD, USA
mdredze@jhu.edu

Abstract

Research on entity linking has considered a broad range of text, including newswire, blogs and web documents in multiple languages. However, the problem of entity linking for spoken language remains unexplored. Spoken language obtained from automatic speech recognition systems poses different types of challenges for entity linking; transcription errors can distort the context, and named entities tend to have high error rates. We propose features to mitigate these errors and evaluate the impact of ASR errors on entity linking using a new corpus of entity linked broadcast news transcripts.

1 Introduction

Entity linking identifies for each textual mention of an entity a corresponding entry contained in a knowledge base, or indicates when no such entry exits (NIL). Numerous studies have explored entity linking in a wide range of domains, including newswire (Milne and Witten, 2008; Mcnamee et al., 2009; McNamee and Dang, 2009; Dredze et al., 2010), blog posts (Ji et al., 2010), web pages (Demartini et al., 2012; Lin et al., 2012), social media (Cassidy et al., 2012; Guo et al., 2013a; Shen et al., 2013; Liu et al., 2013), email (Gao et al., 2014) and multi-lingual documents (Mayfield et al., 2011; McNamee et al., 2011; Wang et al., 2012). A common theme across all these settings requires addressing two difficulties in linking decisions: matching the textual name mention to the form contained in the knowledge base, and using contextual clues to disambiguate similar entities. However, all of these studies have focused on written language, while linking of spoken language remains untested. Yet many intended applications of entity linking, such as supporting search (Hachey et al., 2013) and identifying relevant sources for reports (He et al., 2010; He et al., 2011), linking of spoken language is critical. Search results regularly include audio content (e.g. YouTube) and numerous information sources are audio recordings (e.g. media reports.) An evaluation of entity linking for spoken language can help clarify issues and challenges in this domain.

In addition to the two main challenges discussed above, audio entity linking presents two parallel difficulties that arise from automatic transcription (ASR) of speech. First, the context can be both shorter (than newswire formats) and contain ASR errors, which can make the context of the mention less like supporting material in the knowledge base. Second, named entities are often more difficult to recognize (Huang, 2005; Horlock and King, 2003); they are often out-of-vocabulary and less common overall in training data. This can mislead the name matching techniques on which most entity linking systems depend.

In this paper we consider the task of entity linking for spoken language by evaluating linking on transcripts of broadcast news. We select broadcast news as a comparable domain of spoken language to newswire documents, which have been the focus of considerable research for entity linking (Mcnamee et al., 2009; Ji et al., 2010). We chose this comparable domain to focus on issues introduced because of a transition to spoken language from written, as opposed to issues that arise from a general domain shift associated with conversational speech, an issue that has been previously studied in the shift from written news to written conversations (weblogs, social

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 225–230,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

media, etc.) (Baldwin et al., 2013; Han et al., 2013).

We proceed as follows. We first introduce a new broadcast news dataset annotated for entity linking. We then propose new features based on ASR output to address the two sources of error specific to spoken language: 1) context errors and shortening, 2) name mention transcription errors. We then test our features on the automated output of an ASR system to validate our findings.

2 Entity Linked Spoken Language Data

We created entity linking annotations for HUB4 (Fiscus et al., 1997), a manually-transcribed broadcast news corpus. We used gold named entity annotations from Parada et al. (2011), who manually annotated 9971 utterances with CONLL style (Tjong Kim Sang and De Meulder, 2003) named entities. We selected 2140 person entities and obtained entity linking decisions with regards to the TAC KBP knowledge base (Mcnamee et al., 2009; Ji et al., 2010), which contains 818,741 entries derived from Wikipedia.

Annotations were initially obtained using Amazon Mechanical Turk (Callison-Burch and Dredze, 2010) using the same entity linking annotation scheme as Mayfield et al. (2011). Turkers were asked which of five provided Wikipedia articles matched a person mention highlighted in a displayed utterance. The provided Wikipedia articles were selected based on token overlap between the mention and article title, weighted by TFIDF. In addition to selecting one of the five articles, turkers could select "None of the above", "Not enough information" or "Not a person". We collected one Turker annotation per query and manually verified and corrected each provided label. For mentions that were not matched to an article, we verified that the article was not in the KB, or manually assigned a KB entry otherwise. Because we manually corrected each annotation, mistakes and biases resulting from crowdsourced annotations are not present in this corpus. Of the 2140 annotated mentions, 41% (n=887) were `NIL`, compared to 54.6% in TAC-KBP 2010 (Ji et al., 2010) and 57.1% in TAC-KBP 2009 (Mcnamee et al., 2009). The named entity and linking annotations can be found at `https://github.com/mdredze/speech_ner_entity_linking_data/releases/tag/1.0`.

We divided the 2140 mentions into 60% train ($n = 1283$) and 40% test ($n = 857$.) We ensured that the 1218 unique mention strings were disjoint between the two sets, i.e. no mention in the test data was observed during training.

Each instance is represented by the query (mention string) and document context. Unlike written articles, broadcast news does not indicate where one topic starts and another ends. Therefore, we experimented with different size contexts by including the utterance containing the name mention and up to eight utterances before and after so long as they occurred within 150 seconds of the start of the utterance containing the name mention. We found that five utterances before and after gave the highest average accuracy when using a standard set of features on the reference transcript to perform entity linking; we use this setting in the experiments below.

3 Entity Linking System

For entity linking, we use Slinky (Benton et al., 2014), an entity linking system that uses parallel processing of queries and candidates in a learned cascade to achieve fast and accurate entity linking performance. We use standard features from McNamee et al. (2012) as described in Benton et al. (2014). For a query q composed of a context (multiple utterances) and a named entity string, Slinky first triages (Dredze et al., 2010; McNamee et al., 2012; Guo et al., 2013b) the query to identify a set of candidates C_q in the knowledge base that may correspond to the mention string. Up to 1000 candidates are considered, though its usually much fewer. The system then extracts features based on query and candidate pairs and ranks them using a series of classifiers. The candidates are then ordered by their final scores; the highest ranking candidate is selected as the system's prediction. `NIL` is included as a candidate for every query and is then ranked by the system. For all training settings, we sweep over hyper-parameters using 5-fold cross validation on the training data to find the most accurate system. Reported results are based on the test data.

3.1 Spoken Language Features

ASR transcription errors pose a challenge for standard text processing features, which rely on textual similarity to measure relatedness of both context and entity mention strings. However, ASR errors are not random; incorrectly decoded words may be phonetically similar to the original spoken words. This suggests that similarity can still be captured by considering phonetic similarity.

We experiment with four feature types.

- **Text**: Our baseline system uses features based on the text of the mention string and document. We used the feature set presented by McNamee et al. (2012) and used in Benton et al. (2014), which was the best performing submission in the TAC-KBP 2009 entity linking task. These features include, among others, features that compare the candidate entity name to the mention string as well as the document's terms to those stored in the candidate's description in the KB. These include the dice coefficient, cosine similarity (boolean and weighted), and proportion of candidate tokens in the query document.
- **Phone**: Words in the document, mention string as well as the knowledge base are represented as phone sequences instead of text. We convert all words to phones using a grapheme string to phone (G2P) system.
- **Metaphones**: Two distinct phones can sound similar, yet still appear different when matching phones. Metaphones (Philips, 1990), a more recent version of Soundex, map similar sounding phones to the same representation. We convert the phones used in the previous paragraph to metaphones.
- **Lattice**: Expected word counts of the query document from the ASR lattice. Extracted unigrams are treated as a weighted bag-of-words for the query document. We compute all the features that use the query document's content and weigh them by the term's expectation.

The features in each of the above sets depend on their representation of the text (e.g. text, phone, metaphone, lattice). Additionally, we include the following features in all experiments: *Bias* features that fire for all candidates, only non-`NIL` candidates, and only `NIL` candidates; `NIL` features indicative of being linked to no article in the knowledge base such as the mention string is only 1 or 2 characters/tokens, the number of candidates emitted by the triager; and the *Popularity* (number of Wikipedia in-links) of the candidate.

Triage The above feature sets change the ranking of candidates. We also modified the triage methods that produce C_q based on these new features. For **Text** we used the triage methods of Mcnamee et al. (2009): string similarity based on character/token n-gram overlap, same acronym, exact match, etc. **Phone** triage used the same heuristics but based on phone representations of the mention strings and candidate names. **Metaphone** triage worked as in phone, but used metaphones. When two representations are used we take the union of their candidates.

G2P System For phone features we use a G2P system based on the Sequitur G2P grapheme-phoneme-converter (Bisani and Ney, 2008). We trained the system on version 0.7a of CMUdict[1] (stress omitted from phone sequences, case-insensitive for graphemes), by predicting the phoneme sequence of an English token given its string (G2P). The language model was a 9-gram model with modified Kneser-Ney smoothing applied. For **Phone** features we converted each token to its best phone representation, where each phone, as well as diphthong, is represented by a single character for similarity matching.

4 Experiments

We evaluate reference transcripts and output from two ASR systems run on our dataset (HUB4). We use Kaldi (Povey et al., 2011) trained on the spoken version of the Wall Street Journal corpus (Paul and Baker, 1992).[2] The first system (*mono*) relies on an HMM whose hidden states are context-independent phones. The second system (*tri4b*) uses an HMM that outputs phones dependent on their immediate left and right contexts. These systems respectively achieve 70.6% and 50.7% WER over our training set, where high error rates are likely due to a shift in domain from primarily financial news to the wider

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] Linguistic Data Consortium number LDC94S13A.

	Reference	mono	tri4b
WER	-	71%	51%
Mention WER	-	90%	63%
Mention Exact	-	5%	22%
Text	0.77/0.77	0.44/0.48	0.56/0.55
Phone	0.77/0.79	0.42/0.45	0.47/0.46
Text+Phn	0.81/0.81	0.45/0.49	0.58/0.61*
Text+Phn+Latt	-	0.45/0.49	0.59/0.60*
Metaphone	0.52/0.61	0.43/48	0.52/0.56
Text+Metaphn	0.78/0.78	0.45/0.50	0.59/0.61*
Text+Metaphn+Latt	-	**0.45/0.51**	**0.59/0.63****

Table 1: Performance of different feature sets for reference transcripts and ASR output (mono, tri4b). Results are cross-fold validation/test accuracy. Significance of system test accuracy compared to the Text baseline computed with two-sample proportion test. (*$p < 0.05$, ** $p < 0.01$).

	Text	Phone	Text+Phn
Reference	0.92/0.87	0.95/0.91	0.98/0.97
mono	0.48/0.13	0.50/0.17	0.51/0.19
tri4b	0.68/0.47	0.71/0.52	0.73/0.56

Table 2: Overall/non-NIL triager recall.

news variety in HUB4. These higher error settings test the limits of entity linking in noisy ASR.

To find the ASR-corrupted mention string used for the query q we align the ASR transcript by token-level edit distance – additions and deletions cost 1, while substitutions cost the Jaccard distance between two tokens. This is done at both training and test time, and allows us to evaluate performance of entity linking features without worrying about errors introduced by a named entity recognizer on the transcripts.

Entity Linking System Training The entity linking system relies on a linear Ranking SVM objective (Joachims, 2006), and the optimal slack parameter C was chosen using 5-fold cross validation over the training set (C varied from 1 and 5 $\times 10^{-5 \ldots 3}$). During cross-validation, mention string types were kept disjoint between the train and development folds. Ranking was performed over the (up to) 1000 candidates produced by triage selected from the TAC-KBP 2009/10 KB (Mcnamee et al., 2009; Ji et al., 2010). Using the selected C we trained over the entire training set and evaluated on the test set.

	Text	Phone	Text+Phn	T+P+Latt
Reference	0.82	0.82	0.85	0.85
tri4b	0.76	0.62	0.77	0.78

Table 3: Cross validation accuracy evaluated over only those queries whose correct candidate was output by the triager for both tri4b and reference.

tri4b	Reference
on joseph	Don Joseph
ira magazine	Ira Magaziner
bob defiance	Bob Mathias
gave deforest	Dave Deforest
georgia the books	George W. Bush's
george w. porsche	George W. Bush
louis freer	Louis Freeh
norman monetta	Norman Mineta
edward and	Edward Egan
keith clarke	Nikki Clark

Table 4: Examples of improved linking accuracy.

5 Results

Table 1 reports both the average accuracy for 5-fold cross validation (CV) on train and for the best tuned system from CV on test data. The reference test accuracy is relatively high, but lags behind person entity linking for written language. When accurate transcripts are available, entity linking for spoken language, while harder, achieves just a little behind written language. However, on ASR transcripts, accuracy drops considerably: 0.77 reference to 0.48 (mono, 71% WER) or 0.55 (tri4b, 51% WER). Our features improve accuracy for both ASR systems. Metaphone features do better than Phone features. Lattice do not show significant improvements, likely because they help with context but not mentions (see below.) When combined with text, both metaphone and phone features do similarly.

The majority of our improvement comes from improvements to recall. Table 3 shows the accuracy of queries for which the triager found the correct candidate in *both* the reference transcript and tri4b, providing a consistent set for comparison. For these queries, tri4b is much closer to the results obtained on reference and much higher than the best results in Table 1. This is encouraging, especially given the 50% WER of tri4b; entity linking accuracy is not seriously impacted by noisy transcripts, provided that

the correct candidate is recalled for ranking.

Table 2 shows the recall of the triager on the reference, tri4b and mono transcripts. Reference recall is quite high, while recall for the ASR systems is much lower. Here, our features dramatically improve the recall, giving the ranker an opportunity to correctly score these queries. The challenge of recall is that many of the mention strings are incorrectly recognized. Table 1 shows the WER of the mention string and the number of mention strings for which the recognition is completely correct. Unsurprisingly, error rates for mentions are higher than the overall WER. In short, success on ASR transcripts is primarily dictated by the effectiveness of finding candidates in triage, which is much harder given the low recognition rate. Our features most benefit overall accuracy by improving recall.

Finally, Table 4 provides example of improved recall: mention strings that are incorrectly recognized by the tri4b ASR system leading to linking failures, but are then correctly linked by our improved features. These examples demonstrate the effectiveness of phonetic matching, retrieving the correct "George W. Bush" when the recognizer output "Georgia the Books."

6 Conclusion

We have conducted the first analysis of entity linking for spoken language. Our new features, which rely on phonetic representations of words and expected counts of the lattice for context, improve the accuracy of an entity linker on ASR output. Our analysis reveals that while the linker is not sensitive to large drops in error rates in the context, it is highly sensitive to error rates in mention strings, due to a drop in triage recall. Our features improve the overall accuracy by improving the recall of the triager. Future work should focus on additional methods for identifying relevant KB candidates given inaccurate transcriptions of mention strings.

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364.

Adrian Benton, Jay Deyoung, Adam Teichert, Mark Dredze, Benjamin Van Durme, Stephen Mayhew, and Max Thomas. 2014. Faster (and better) entity linking with cascades. In *NIPS Workshop on Automated Knowledge Base Construction*.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. In *Speech Communication*, volume 50, pages 434–451.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Workshop on Creating Speech and Language Data With Mechanical Turk at NAACL-HLT*.

Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, and Hongzhao Huang. 2012. Analysis and enhancement of wikification for microblogs with context expansion. In *COLING*, pages 441–456.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA. ACM.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *COLING*.

Jonathan Fiscus, John Garofolo, Mark Przybocki, William Fisher, and David Pallett. 1997. 1997 english broadcast news speech (hub4). In *LDC98S71*.

Ning Gao, Douglas Oard, and Mark Dredze. 2014. A test collection for email entity linking. In *NIPS Workshop on Automated Knowledge Base Construction*.

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013a. To link or not to link? a study on end-to-end tweet entity linking. In *NAACL*.

Yuhang Guo, Bing Qin, Yuqin Li, Ting Liu, and Sheng Li. 2013b. Improving candidate generation for entity linking. In *Natural Language Processing and Information Systems*, pages 225–236. Springer.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with wikipedia. *Artificial Intelligence*, 194(0):130 – 150. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27, February.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference*

on World Wide Web, WWW '10, pages 421–430, New York, NY, USA. ACM.

Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 755–764, New York, NY, USA. ACM.

James Horlock and Simon King. 2003. Named entity extraction from word lattices. *Eurospeech*.

Fei Huang. 2005. *Multilingual Named Entity Extraction and Translation from Text and Speech*. Ph.D. thesis, Carnegie Mellon University.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. *Third Text Analysis Conference (TAC 2010)*.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Thomas Lin, Oren Etzioni, et al. 2012. Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 84–88. Association for Computational Linguistics.

Xiaohua Liu, Yitong Li, Haocheng Wu, Ming Zhou, Furu Wei, and Yi Lu. 2013. Entity linking for tweets. In *ACL*.

James Mayfield, Dawn Lawrie, Paul McNamee, and DouglasW. Oard. 2011. Building a cross-language entity linking collection in twenty-one languages. In Pamela Forner, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas, and Marteen de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation*, volume 6941 of *Lecture Notes in Computer Science*, pages 3–13. Springer Berlin Heidelberg.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *TAC*.

Paul Mcnamee, Mark Dredze, Adam Gerber, Nikesh Garera, Tim Finin, James Mayfield, Christine Piatko, Delip Rao, David Yarowsky, and Markus Dreyer. 2009. Hltcoe approaches to knowledge base population at tac 2009. In *Text Analysis Conference (TAC*.

Paul McNamee, James Mayfield, Douglas W. Oard, Tan Xu, Ke Wu, Veselin Stoyanov, and David Doerman. 2011. Cross-language entity linking in maryland during a hurricane. In *TAC*.

Paul McNamee, Veselin Stoyanov, James Mayfield, Tim Finin, Tim Oates, Tan Xu, Douglas Oard, and Dawn Lawrie. 2012. HLTCOE participation at TAC 2012: Entity linking and cold start knowledge base construction. In *TAC*.

David N. Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *CIKM*, pages 509–518.

Carolina Parada, Mark Dredze, and Frederick Jelinek. 2011. Oov sensitive named-entity recognition in speech. In *International Speech Communication Association (INTERSPEECH)*.

Douglas Paul and Janet Baker. 1992. The design for the wall street journal-based csr corpus. In *DARPA Speech and Language Workshop*. Morgan Kaufmann Publishers.

Lawrence Philips. 1990. Hanging on the metaphone. *Computer Language*, 7(12 (December)).

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2013. Linking named entities in tweets with knowledge base via user interest modeling. In *KDD*.

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 seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 459–468, New York, NY, USA. ACM.