Improving Named Entity Recognition with Prosodic Features

Denys Katerenchuk[1], Andrew Rosenberg[1,2]

[1]CUNY Graduate Center, New York, USA
[2]CUNY Queens College, New York, USA
dkaterenchuk@gc.cuny.edu, andrew@cs.qc.cuny.edu

Abstract

In natural language processing (NLP) the problem of named entity (NE) recognition in speech is well known, yet remains a challenge where performance is dependent on automatic speech recognition (ASR) system error rates. NEs are often foreign or out-of-vocabulary (OOV) words, leaving conventional ASR systems unable to recognize them. In our research, we improve a CRF-based NE recognition system by incorporating two styles of prosodic features, hypothesized ToBI labels and unsupervised clusters of acoustic features. ToBI-based features improve NE recognition by 6% absolute (F1:0.39 v.s. F1: 0.45) on automatically recognized spontaneous speech from ACE'05.

1. Introduction

Named Entity recognition is one of the core problems in NLP. A wealth of tasks including summarization, translation, question generation and a wealth of information extraction tasks all rely on accurate NE detection. NEs are critically important to understanding the content of a document.

In this work, we identify NEs that define geo-political entities (GPE), persons (PER), and organizations (ORG). Inability to recognize these words directly reflects on natural language understanding. Many NEs are foreign or previously unseen out-of-vocabulary (OOV) words. This can also makes the task challenging for humans. Consider the following excerpt from a Wall Street Journal article: *Mr.* ***Ahmadi*** *declined to discuss whether any letter or other reassurance was provided by the* ***U.S.*** *on Wednesday, saying more details would be forthcoming on Thursday morning, when Mr.* ***Karzai*** *is scheduled to make an opening speech at the* ***Loya Jirga.***

In this paragraph, NEs are bold. A person asked to identify NEs would have no problem identifying those words that follow each "Mr." as names, regardless of familiarity. Although abbreviated, it is safe to say that this person would recognize "U.S." as a NE. The last phrase, however, is more likely to bring uncertainty. Without knowing the Afghani and Pakistani cultures, this could be identified as a place name. Inability to recognize "Loya Jirga" as an organization brings ambiguity to the entire sentence.

Since NEs are crucial for language understanding, many systems, including those used in the Automatic Content Extraction (ACE), GALE and Knowledge Base Population (KBP) programs, use NE recognition (NER) in their pipelines. However, one domain where NER performance remains problematically low is on speech data. NE recognition depends on an automatic speech recognition (ASR) system that will introduce error. Moreover, speech is less formal and more idiosyncratic across speakers. Speakers are less likely to use strict syntactic constructions, frequently using unconventional expressions. Disfluencies are common in speech. From a system building perspective, there is also less available transcribed and annotated speech data than there are text documents.

People, of course, have no problem detecting NEs in speech. Consider the possibility that the person may be listening to someone read the previous paragraph, rather than reading it. Even if a listener has never heard or knows how to spell "Ahmadi", "Karzai" and "Loya Jirga", they have a reasonably good chance at recognizing that "Ahmadi" and "Karzai" are people, and that "Loya Jirga" is either a place or organization. In addition to using the contextual cues described previously, people make use of acoustic cues to understand the structure and focus of spoken language. We hypothesize that these acoustic cues, encoding information about *how* the words were said, can be used in concert with *what* words were spoken to improve NER, particularly when operating on moderate to high word-error rate (WER) ASR recognition hypotheses. We explore two prosodic representations. In one, we hypothesize ToBI labels using the AuToBI toolkit. The other representation generates discrete prosodic labels by clustering low-level acoustic/prosodic features.

The rest of the paper is structured as follows. In Section 2, we describe related work on NER in speech. We describe the major components of our NER system in Section 3 including the ASR systems and hypothesis structures we explore (Sections 3.1), the two styles of prosodic analysis (Sections 3.3 and 3.4) and details of the NER system proper (Section 3.2). We then describe the data for training ASR and NER systems (Section 4) and results from NER experiments (Section 5). We conclude in Section 6.

2. Related Work on NER in Speech

Named entity detection in speech is not a new problem. An early investigation measures the impact of case information and ASR errors on task performance [1]. The simplest approach involves applying a NER system that is trained on annotated text to ASR transcripts. However, this tends to work quite poorly. A somewhat better approach involves training the system on manual transcriptions of similar speech data [2]. A variety of approaches have been developed to improve recognition and detection of named entities. [3] focus their attention on identifying which words in an utterance are out-of-vocabulary (OOV) to improve NER on speech data. [4] used metadata about the recording location to adapt the ASR vocabulary to include more relevant names and locations. [5] applied WFSTs to recognize named entities in consensus nets finding a 6-10% absolute improvement by looking at hypotheses other than the ASR 1-best.

Another approach uses acoustic information to punctuate ASR output [6]. The idea of this approach is that punctuation carries structural and contextual information that is valuable for NE detection [7, 8]. This information is explicit in text, but

Copyright © 2014 ISCA

cannot be found in ASR results. An approach as simple as using pause information as a proxy for punctuation has been shown to improve NE detection in speech.

Hakkani Tur et al. use prosodic features for binary NE classification in speech [9]. They use f0 pattern, pitch range and boundaries, as input to a decision tree classification to recognize names. In this approach prosody alone was used to effectively recognize named entities, but didn't give any improvement to NER in combination with lexical features.

The study that is most similar to this paper was done by [10]. This work investigates the use of prosody in detecting names in template sentences. Each word is classified as a Name or not using prosodic features. This prediction is combined with a lexical HMM name tagger. While the performance they report is quite high, finding 80% detection accuracy, the task is quite different than NE recognition on news data or spontaneous speech. They worked on isolated sentences and each sentences contained the first mention of a name. First mentions often lead to increased prosodic emphasis. In addition to this, only recognition of proper names is required. No geopolitical entities or organization names were present. This makes both the ASR and NE recognition easier in this task than on the ACE material (Section 4).

3. Methods

In this section, we describe the speech recognizer (Section 3.1) and named entity recognizer (Section 3.2) used in our experiments. We explain the two types of prosodic features in Sections 3.3 and 3.4.

3.1. Automatic Speech Recognition Systems

For the experiments in this paper, we use the KALDI ASR Toolkit developed by [11]. Using standard training recipes for the ASR, we build two different acoustic models trained on WSJ data [12]. Both models are trained using 12 MFCC coefficients + energy, delta and double deltas.

The first model is a triphone GMM model trained with ML criteria, without discriminative training or speaker adaptation. This is far from the state-of-the-art in speech recognition, even within KALDI package. We use this recognizer to have a high WER operating point that is consistent with using an ASR system as a blackbox, where there may be significant inconsistencies between training and testing channels, speakers and recording conditions. The WER on WSJ data, ASR training corpus, is 18.3%. The second model is a subspace Gaussian mixture model (SGMM) with fMLLR transforms [13]. This is a more effective recognizer. The WER using this model is 13.4%. These two models allow us to assess the relative impact of prosody in improving NE recognition at different WER levels. Regardless of which ASR system is used, we decode using Minimal Bayes Risk (MBR) decoding [14] to generate a 1-best hypotheses.

3.2. Named Entity Recognizer

For NER training, we use ASR decoded speech data. We align annotated transcripts to ASR 1-best hypotheses. Since we know that the hypotheses will contain errors, we apply a minimum edit distance approach to align the tags. This method, however, introduces the problem of word deletions, where a word in the manual transcript does not align to any word in the hypothesis, and substitutions, where the content of a NE tag in the hypotheses may be different from the reference transcript. We remove annotations on deleted words, and treat everything else as a true annotation, even where a NE tag may be misaligned or misrecognized.

We generate evaluation annotations of the test data using the same process as described for the training data. This means that some NE tags contain misrecognitions, and in the case of deletions, the true number of NEs may be higher than the number used in the evaluation. Considering the impact of this decision, correctly identifying a NE annotation does not mean that the ASR system correctly recognize the name itself. There is previous research concerning the detection and recovery of OOV spellings (e.g. [15]); we treat this as a necessary post-processing step to this work.

We use the English NE tagger described in [16]. The tagger is based on standard linear-chain Conditional Random Fields (CRFs) [17] with a rich set of lexical, syntactic and dictionary-based features. The model re-casts the NE task as a token-based sequential labeling, where each token is assigned a label from BIO tag set to indicate whether the token is Beginning, Inside, and Outside of a name. We refer the reader to [16] for detailed description of the baseline features. Some of the features cannot be extracted from ASR hypotheses, including those based on capitalization and punctuation. In addition, syntactic features including POS tags are vulnerable to ASR errors.

3.3. AuToBI-based Prosodic features

One of the two prosodic representations we use in this work is based on automatically hypothesized ToBI [18] labels generated by the AuToBI tool.

The ToBI Standard describes Standard American English (SAE) intonation in terms of **break indices** describing the degree of disjuncture between consecutive words, and **tones** which are associated with phrase boundaries and pitch accents. Pitch accented words are prominent from the surrounding utterance. Five types of pitch accents – pitch movements that correspond to perceived prominence of an associated word – are defined in the standard: H*, L*, L+H*, L*+H, H+!H*. In addition to these five, high tones (H) can be produced in a compressed pitch range indicated as !H.

Two levels of prosodic phrasing are defined, intonational phrases boundaries are defined by the highest degree of disjuncture, and are often associated with silence. Each intonational phrase is comprised of one or more weak or intermediate phrases. Each intermediate phrase has an associated phrase accent, describing the pitch movement between the ultimate pitch accent and the phrase boundary. Phrase accents can have High (H-), downstepped High (!H-) or low (L-) tones. Intonational phrase boundaries have an additional boundary tone, to describe a final pitch movement. These can be high (H%) or low (L%). Since each intonational phrase boundary also terminates an intermediate phrase, intonational phrase boundaries have associated phrase accents *and* boundary tones. Each intermediate phrase must contain at least one pitch accent.

AuToBI [19, 20] is a system to automatically hypothesize the presence and type of prosodic events that are present in a spoken utterance. Automatic generation of ToBI labels consists of six tasks: 1) detection of pitch accents, 2) classification of pitch accent types, 3) detection of intonational phrase boundaries, 4) detection of intermediate phrase boundaries, 5) classification of intonational phrase ending tones, and 6) classification of intermediate phrase ending tones. The system optionally accepts an input segmentation of the signal into words. Using AuToBI, we generate 6 features to incorporate into the

NE recognition feature vector, corresponding to the 6 tasks. This is not exactly identical to a hypothesized ToBI labeling of the input speech; detection and classification predicted as separate features. When generating ToBI labeling, if pitch accent is not detected, for example, prediction of pitch accent is ignored *type*. In this context, the feature vector always includes prosodic event *classification* hypotheses, regardless of whether the detectors identify an event. While less motivated by prosodic theory, the inclusion of these classification predictions provides information about the prosodic contour shape. We find that including these improves NE recognition on some tokens (cf. Section 5). As we are using a CRF for NER, we use the predicted class as a categorical variable.

AuToBI Performance: The AuToBI models used in this work are trained on pooled data from the Boston University Radio News Corpus (2.35 hours of broadcast news-like speech from 6 speakers) [21], Boston Directions Corpus (1.83 hours of spontaneous and read speech from 4 speakers) [22] and Columbia Games Corpus (9 hours of spontaneous dialog speech from 12 speakers) [23]. AuToBI generalizes fairly well to new data; cross-corpus performance is generally 2%-4% points worse than speaker-independent, within-corpus performance on all 6 tasks [24]. The expected F_1 performance of detecting pitch accents (Logistic Regression) is 84.53%, intonational phrases (Logistic Regression) is 74.67% and intermediate phrases (Logistic Regression) 40.05%. The expected Average Recall in classifying pitch accents (AdaBoost) is 18.41%, phrase accents (Random Forests) is 43.96% and phrase accent / boundary tones (Random Forests) is 30.67%. These represent state of the art results on cross-corpus automatic ToBI labeling.

AuToBI Features: A thorough description of the features used by AuToBI can be found elsewhere [19, 20], here we include a high-level description of the types of features used. The general framework is to extract "contours", time-aligned information streams drawn from acoustic/prosodic qualities. Features are extracted from these contours, by applying some function to contours drawn from regions of analysis based on the input segmentation – though many of these regions incorporate surrounding context. The contours include pitch (log Hz), intensity, and spectral tilt, normalized forms of each, and deltas. Also combined contours constructed to capture the interaction of these information streams. From these a number of simple aggregations (mean, max, min, standard deviation, z-score of the max relative to the region) is extracted and some more complicated features that attempt to capture contour shape (TILT features, Center of Gravity, Area under the Contour, Quantized Contour Model posteriors, isotonic regression-based posteriors). To capture contextual information extracted features are normalized by their surrounding context and take difference between features extracted from one word to the following word. While the six AuToBI tasks use distinct feature sets, the union comprises 310 features.

3.4. Clustering

ToBI derived features have had some, but limited, success in application to spoken language processing tasks. A more common approach is, so called, direct-modeling of prosody [25]. In this approach, low-level acoustic/prosodic features are directly added to a feature vector.

The clustering-based prosodic representation described in this section is more appropriately considered to be a direct modeling approach to prosodic analysis. The approach we take here is to extract the union of acoustic/prosodic features used in all AuToBI prediction tasks for each word. We first whiten this data, standardizing each feature to unit-variance. Note that this whitening statistic is calculated over training data only. We then cluster this data using k-means at k in the range [2,10]. This is essentially k-means vector quantization generating a codebook of sizes 2-10. We fit these clusters only on the NE recognition training data. The whitening, cluster training and assignment are performed using scikit-learn [26]. Less than 1% of the time, AuToBI generates a feature vector with a missing value. This can happen when a word is completely unvoiced – no pitch contour is extracted within a region – or when the a standard deviation of some feature is zero or undefined. We assign those words which have a feature vector with a missing value to a unique cluster and remove them from the whitening/clustering process. Therefore, while we run the clustering with k between 2 and 10, the resulting dimensionality of the 9 features are 3 to 11 due to this extra MISSING cluster. During both training and testing, we assign a single cluster to each word. Each of the 9 discrete cluster assignments are appended to the feature vector.

4. Data

We use two different corpora in our experiments. The ASR system is trained on English WSJ corpus [12]. This is a clear, professionally read speech audio recording with minimal disfluencies. In total, this corpus includes 78 hours of audio and vocabulary size of $\sim$ 20,000 words. This is a small amount of training to train a state-of-the-art ASR system. Commercial systems are trained on orders of magnitude more speech data. This recognizer represents what can be expected from a system that can be deployed rapidly and/or with limited resources. For NER pipeline experiments, we use the ACE'05 [27] corpus. This is CNN broadcast news recordings in total of 5 hours of data collected from March to June 2003. This data includes a combination of read and spontaneous speech with fairly low rate of disfluency, though the broadcasts include phone conversations, music and some background noise. The transcripts are human annotated and contain NEs, such as person (PER), geo-politic entity (GPE), and organization (ORG). We use the first corpus, WSJ, only to train ASR system and the second, ACE'05, though out the pipeline. ACE'05 files are split into 75%/25% training and evaluation sets.

5. Results

To show impact prosodic features on different levels of ASR quality, we conduct experiments based on two WSJ trained ASR models: triphome model, to test speech recognizers with higher WER, and SGMM, with lower WER. We use WCN decoding to decrease WER in 1-best hypotheses. Work by [1] shows a direct impact of WER on NER performance. WER results of ACE'05 are shown in Table 1.

	Triphone model	SGMM model
WER	67.37%	49.13%
NE-WER	72.55%	59.42%

Table 1: ASR model WER and Named Entity WER on ACE'05

One of the challenges to NER in speech is the impact of OOV words and spelling inconsistencies. Even a perfect ASR system misrecognizes NEs if they are OOV. ASR shows worse performance on NEs compare to regular words. We find an error rate on NE tokens is between 5 and 10 points higher than the

True transcription	NE tagger	NE tagger + Prosody
secretary-general did not answer questions on Iraq	secretary general did not answer questions on a rock	secretary general did not answer questions on <GPE> a </GPE> rock
battling for control of the bridges in the southern city of Nasiriyah	battling for control the bridges in the southern city of non sir re f	battling for control the bridges in the southern city of non <GPE> sir </GPE> re f
from Iraqi paramilitary groups today	from the rocket hair and military groups today	from the <GPE> rocket </GPE> hair and military groups today

Table 2: NE Tagger performance on ASR hypotheses

overall WER (cf. Table 1) and 14% of NE types are OOV.

We conduct our experiments on both, triphone and SGMM, based decodings. NE tagger is first trained on ASR output alone, and then on combination of ASR output and both types of prosodic features. In total, we create four models: 1) Text based features - baseline, 2) Prosodic clusters and text features, 3) AuToBI features and text features, and 4) Prosodic clusters, AuToBI features and text features. The triphone and SGMM results can be found in Table 3.

System	Triphone	SGMM
Baseline	27.65	39.38
Clusters	30.57	39.94
AuToBI	**30.75**	**45.02**
Clusters+AuToBI	29.10	44.34

Table 3: NE tagger results on triphone and SGMM recognizers

On the triphone recognizer prosodic clusters outperform base line by 2.92% and AuToBI features by 3.1% absolute. The AuToBI features represent a relative improvement of over 11%. In combination, however, while still outperforming the baseline by 1.45%, these features show lower performance then either alone. NE tagger with acoustic features also performs well on SGMM recognizer output. The WER is improved by 18.24% in the SGMM system, and the baseline NER F_1 increases by 11.73%. Makhoul et al. finds that a relative reduction in WER should lead to an relative improvement of approximately half the magnitude to NER [8], however, we find a smaller increase. On the SGMM hypotheses, we again find prosody to yield substantial improvements to NER. Cluster based features show a small, 0.52%, improvement. While AuToBI features alone raise the results by 5.64% absolute, a relative improvement of 14%.

AuToBI hypotheses improve NER performance in a way that is largely invariant to speech recognition quality. There is an expectation that the impact of prosody other non-lexical information would have a greater impact when recognition performance is lower. The utility of this information is that it is robust to recognition errors and provides an additional information stream to the NER system. As the reliability of ASR transcripts improve, the improvement offered by this additional information stream may diminish. This is not observed at the WER points that we have examined. Despite reducing the WER by 18.24%, the improvement provided by including prosodic information as AuToBI features to WER *increases*.

There are a few possible explanations to what information hypothesized ToBI labels are providing to assist NER. Pitch accents encode prosodic prominence; named entities are typically focused in speech, making them likely to be prominent. Additionally, some researchers have investigated the impact of inserting punctuation into ASR transcripts as a route to improving NER [6, 7, 8], frequently using acoustic/prosodic information to identify sentence boundaries and commas. Prosodic phrasing is related, though not isomorphic, to punctuation in speech. Sentence boundaries are almost always intonational phrase boundaries. Speakers also often include a phrase boundary at commas. Identification of prosodic phrases and their associated intonation may be providing similar information to the NER as punctuation does.

We find that prosodic cues help identify NEs in misrecognized text. Table 2 provides example of true transcripts and ASR hypotheses annotated by our NE tagging using baseline and AuToBI features. In all the cases, ASR misrecognized NEs. Despite this fact, prosodic features help to identify the NEs that were missed by NE tagger alone.

6. Conclusion and Future Work

In this work we address the problem of recognizing named entities in speech. We focus specifically on identifying prosodic features which are able to improve named entity recognition (NER) in the context of high and medium error rate (67.37-49.13% WER) speech recognition output.

We find that predictions of ToBI-style prosodic events by AuToBI provide substantial and consistent improvements to NER at both WER conditions; 14% relative (5.64% absolute F_1 gain) in the lower WER case, 11% relative (3.1% absolute) in the higher WER case. Incorporating acoustic information through k-means clusters (or VQ codebook) features yields similar, improvements in the high WER case, but these nearly vanish in the lower WER case. Combination of these two types of prosodic features consistently, if modestly, reduces performance. One explanation for this improvement is that the prosodic events that are hypothesized by AuToBI are serving as a proxy for punctuation. Comparing the performance of this approach to that obtained by punctuation prediction will be a useful next step. These may represent redundant or complementary information. One limitation of this work is that while we are able to identify NEs, these are frequently misrecognized; the error rate is higher on NEs than other terms. An important extension will be the spelling recovery and OOV detection to make the recognized NEs more useful to downstream tasks. In doing so we will extend this approach to other information extraction tasks.

We are continuing to improve our ASR system, as we do this, we will evaluate this work on lower WER conditions to examine whether or not the improvement due to prosody continues to be present.

7. Acknowledgments

This material is based on research sponsored by DARPA under agreement number FA8750-13- 2-0041. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

8. References

[1] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," in *in Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 287–292.

[2] B. Favre, F. Béchet, and P. Nocéra, "Robust named entity extraction from large spoken archives," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 491–498. [Online]. Available: http://dx.doi.org/10.3115/1220575.1220637

[3] C. Parada, M. Dredze, and F. Jelinek, "Oov sensitive named-entity recognition in speech." in *INTERSPEECH*. ISCA, 2011, pp. 2085–2088.

[4] B. Ramabhadran, O. Siohan, and G. Zweig, "Use of metadata to improve recognition of spontaneous speech and named entities." in *INTERSPEECH*. ISCA, 2004.

[5] D. Z. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tür, "Beyond asr 1-best: Using word confusion networks in spoken language understanding." *Computer Speech and Language*, vol. 20, no. 4, pp. 495–514, 2006.

[6] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tür, and M. Ostendorf, "Punctuating speech for information extraction." in *ICASSP*. IEEE, 2008, pp. 5013–5016.

[7] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tür, M. P. Harper, M. Ostendorf, and W. Wang, "Impact of automatic comma prediction on pos/name tagging of speech." in *SLT*, M. Gilbert and H. Ney, Eds. IEEE, 2006, pp. 58–61.

[8] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. A. Ramshaw, D. Stallard, R. M. Schwartz, and B. Xiang, "The effects of speech recognition and punctuation on information extraction performance." in *INTERSPEECH*. ISCA, 2005, pp. 57–60.

[9] D. Z. Hakkani-Tür, G. Tür, A. Stolcke, and E. Shriberg, "Combining words and prosody for information extraction from speech," in *EUROSPEECH*, 1999.

[10] V. Rangarajan and S. S. Narayanan, "Detection of non-native named entities using prosodic features for improved speech recognition and translation," in *ISCA Multiling Workshop*, 2006.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[12] J. Garofalo et al., "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[13] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model - a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523081000063X

[14] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Comput. Speech Lang.*, vol. 25, no. 4, pp. 802–828, Oct. 2011.

[15] L. Qin, M. Sun, and A. I. Rudnicky, "Oov detection and recovery using hybrid models with different fragments." in *INTERSPEECH*. ISCA, 2011, pp. 1913–1916.

[16] Q. Li, H. Li, H. Ji, W. Wang, J. Zheng, and F. Huang, "Joint bilingual name tagging for parallel corpora." in *CIKM*, X. wen Chen, G. Lebanon, H. Wang, and M. J. Zaki, Eds. ACM, 2012, pp. 1727–1731.

[17] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[18] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[19] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[20] ——, "Modeling intensity contours and the interaction between pitch and intensity to improve automatic prosodic event detection and classification," in *Interspeech*, 2012.

[21] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep. ECS-95-001, March 1995.

[22] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[23] A. Gravano and J. Hirschberg, "Turn-yielding cues in task-oriented dialogue," in *SigDial*, 2009.

[24] A. Rosenberg, "Classification of prosodic events using quantized contour modeling," in *HLT-NAACL*, 2010.

[25] E. Shriberg and A. Stolcke, "Direct modeling of prosody: an overview of applications in automatic speech processing," in *Speech Prosody*, 2004, pp. 575–582.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] I. Mani, J. Hitzeman, J. Richer, and D. Harris, "Ace 2005 english spatialml annotations," *Linguistic Data Consortium, Philadelphia*, 2008.