Improving coreference resolution with automatically predicted prosodic information

Ina Rösiger*, **Sabrina Stehwien***, **Arndt Riester**, **Ngoc Thang Vu**
Institute for Natural Language Processing
University of Stuttgart, Germany
{roesigia,stehwisa,arndt,thangvu}@ims.uni-stuttgart.de

Abstract

Adding manually annotated prosodic information, specifically pitch accents and phrasing, to the typical text-based feature set for coreference resolution has previously been shown to have a positive effect on German data. Practical applications on spoken language, however, would rely on automatically predicted prosodic information. In this paper we predict pitch accents (and phrase boundaries) using a convolutional neural network (CNN) model from acoustic features extracted from the speech signal. After an assessment of the quality of these automatic prosodic annotations, we show that they also significantly improve coreference resolution.

1 Introduction

Noun phrase coreference resolution is the task of grouping noun phrases (NPs) together that refer to the same discourse entity in a text or dialogue. In Example (1), taken from Umbach (2002), the question for the coreference resolver, besides linking the anaphoric pronoun *he* back to *John*, is to decide whether *an old cottage* and *the shed* refer to the same entity.

(1) {John}$_1$ has {an old cottage}$_2$.
Last year {he}$_1$ reconstructed {the shed}$_?$.

Coreference resolution is an active NLP research area, with its own track at most NLP conferences and several shared tasks such as the CoNLL or SemEval shared tasks (Pradhan et al., 2012; Recasens et al., 2010) or the CORBON shared task 2017[1]. Almost all work is based on text, although there exist a few systems for pronoun resolution in transcripts of spoken text (Strube and Müller, 2003; Tetreault and Allen, 2004). It has been shown that there are differences between written and spoken text that lead to a drop in performance when coreference resolution systems developed for written text are applied on spoken text (Amoia et al., 2012). For this reason, it may help to use additional information available from the speech signal, for example prosody.

In West-Germanic languages, such as English and German, there is a tendency for coreferent items, i.e. entities that have already been introduced into the discourse (their information status is *given*), to be deaccented, as the speaker assumes the entity to be salient in the listener's discourse model (cf. Terken and Hirschberg (1994); Baumann and Riester (2013); Baumann and Roth (2014)). We can make use of this fact by providing prosodic information to the coreference resolver. Example (2), this time marked with prominence information, shows that prominence can help us resolve cases where the transcription is potentially ambiguous[2].

(2) {John}$_1$ has {an old cottage}$_2$.
a. Last year {he}$_1$ reconstructed {the SHED}$_3$.
b. Last year {he}$_1$ reconSTRUCted **the shed**$_2$.

The pitch accent on *shed* in (2a) leads to the interpretation that *the shed* and *the cottage* refer to different entities, where the shed is a part of the cottage (they are in a bridging relation). In contrast, in (2b), *the shed* is deaccented, which suggests that *the shed* and *the cottage* corefer.

A pilot study by Rösiger and Riester (2015) has

*The two first authors contributed equally to this work.
[1] http://corbon.nlp.ipipan.waw.pl/

[2] The anaphor under consideration is typed in boldface, its antecedent is underlined. Accented syllables are capitalised.

Proceedings of the First Workshop on Speech-Centric Natural Language Processing, pages 78–83
Copenhagen, Denmark, September 7–11, 2017.

shown that enhancing the text-based feature set for a coreference resolver, consisting of e.g. automatic part-of-speech (POS) tags and syntactic information, with pitch accents and prosodic phrasing information helps to improve coreference resolution of German spoken text. The prosodic labels used in the experiments were annotated manually, which is not only expensive but not applicable in an automatic pipeline setup. In our paper, we present an experiment in which we replicate the main results from the pilot study by annotating the prosodic information automatically, thus omitting any manual annotations from the feature set. We show that adding prosodic information significantly helps in all of our experiments.

2 Prosodic features for coreference resolution

Similar to the pilot study, we make use of *pitch accents* and *prosodic phrasing*. We predict the presence of a pitch accent[3] and use phrase boundaries to derive nuclear accents, which are taken to be the last (and perceptually most prominent) accent in an intonation phrase. This paper tests whether previously reported tendencies for manual labels are also observable for automatic labels, namely:

Short NPs Since long, complex NPs almost always have at least one pitch accent, the presence and the absence of a pitch accent is more helpful for shorter phrases.

Long NPs For long, complex NPs, we look for nuclear accents that indicate the phrase's overall prominence. If the NP contains a nuclear accent, it is assumed to be less likely to take part in coreference chains.

We test the following features that have proven beneficial in the pilot study. These features are derived for each NP.

Pitch accent presence focuses on the presence of a pitch accent, disregarding its type. If one accent is present in the NP, this boolean feature gets assigned the value *true*, and *false* otherwise.

Nuclear accent presence is a boolean feature comparable to pitch accent presence. It gets assigned the value *true* if there is a nuclear accent present in the NP, *false* otherwise.

[3]We do not predict the pitch accent type (e.g. fall H*L or rise L*H) as this distinction was not helpful in the pilot study and is generally more difficult to model.

3 Data

To ensure comparability, we use the same dataset as in the pilot study, namely the DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014), a German radio news corpus annotated with both manual coreference and manual prosody labels. We adopt the official train, test and development split[4] designed for research on coreference resolution. The recorded news broadcasts in the DIRNDL-anaphora corpus were spoken by 13 male and 7 female speakers, in total roughly 5 hours of speech. The prosodic annotations follow the GToBI(S) standard for pitch accent types and boundary tones and are described in Björkelund et al. (2014). In this study we make use of two class labels of prosodic events: all accent types (marked by the standard ToBI *) grouped into a single class (pitch accent presence) and the same for intonational phrase boundaries (marked by %).

4 Automatic prosodic information

In this section we describe the prosodic event detector used in this work. It is a binary classifier that is trained separately for either pitch accents or phrase boundaries and predicts for each word, whether it carries the respective prosodic event.

4.1 Model

We apply a convolutional neural network (CNN) model, illustrated in Figure 1. The input to the CNN is a matrix spanning the current word and its right and left context word. The input matrix is a frame-based representation of the speech signal. The signal is divided into overlapping frames for each 20 ms with a 10 ms shift and are represented by a 6-dimensional feature vector for each frame.

We use acoustic features as well as position indicator features following Stehwien and Vu (2017) that are simple and fast to obtain. The acoustic features were extracted from the speech signal using the OpenSMILE toolkit (Eyben et al., 2013). The feature set consists of 5 features that comprise acoustic correlates of prominence: smoothed fundamental frequency (f0), RMS energy, PCM loudness, voicing probability and Harmonics-to-Noise Ratio. The position indicator feature is appended as an extra feature to the input matrices (see Figure 1) and aids the modelling of the acoustic con-

[4]http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.en.html

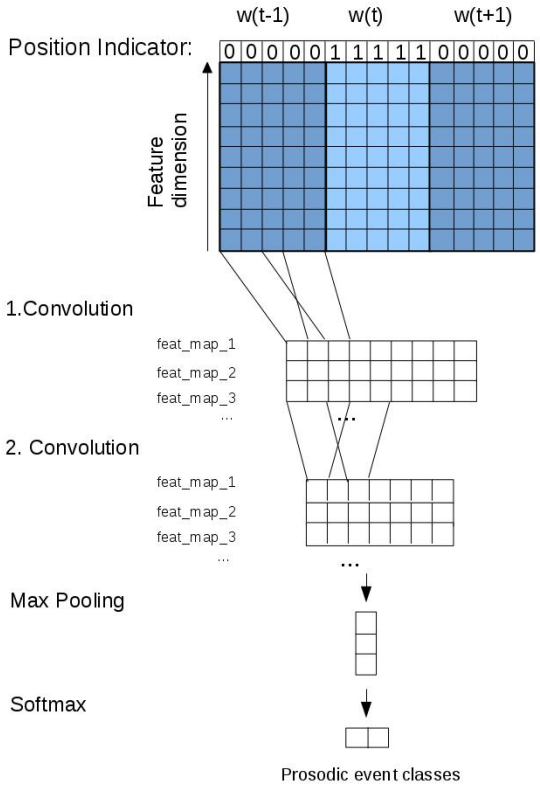

Figure 1: CNN for prosodic event recognition with an input window of 3 successive words and position indicating features.

text by indicating which frames belong to the current word or the neighbouring words.

We apply two convolution layers in order to expand the input information and then use max pooling to find the most salient features. In the first convolution layer we ensure that the filters always span all feature dimensions. All resulting feature maps are concatenated to one feature vector which is fed into the two-unit softmax layer.

4.2 Predicting prosodic labels on DIRNDL

We predict prosodic events for the whole DIRNDL subcorpus used in this paper. To simulate an application setting, we train the CNN model on a different dataset. Since the acoustic correlates of prosodic events as well as the connection between sentence prosody and information status exploited in this paper are similar in English and German, we train the prosodic event detector on English data and apply the model to the German DIRNDL corpus[5]. The data used to train the model is a 2.5 hour subset of the Boston University Radio News Corpus (Ostendorf et al., 1995) that contains speech from 3 female and 2 male speakers and that includes manually labelled pitch accents and intonational phrase boundary tones. Hence, both corpora consist of read speech by radio news anchors. The prediction accuracy on the DIRNDL anaphora corpus is 81.9% for pitch accents and 85.5% for intonational phrase boundary tones[6]. The speaker-independent performance of this model on the Boston dataset is 83.5% accuracy for pitch accent detection and 89% for phrase boundary detection. We conclude that the prosodic event detector generalises well to the DIRNDL dataset and the obtained accuracies are appropriate for our experiments.

5 Coreference resolution

In this section, we describe the coreference resolver used in our experiments and how it was applied to create the baseline system using only automatic annotations.

5.1 IMS HotCoref DE

The IMS HotCoref DE coreference resolver is a state-of-the-art tool for German[7] (Rösiger and Kuhn, 2016). It is data-driven, i.e. it learns from annotated data with the help of pre-defined features using a structured perceptron that models coreference within a document as a directed tree. This way, it can exploit the tree structure to create non-local features (features that go beyond a pair of NPs). The standard features are text-based and consist mainly of string matching, part of speech, constituent parses, morphological information and combinations thereof.

5.2 Coreference resolution using automatic preprocessing

As we aim at coreference resolution applicable to new texts, all annotations used to create the text-based features are automatically predicted using NLP tools. It is frequently observed that the performance drops when the feature set is derived in this manner compared to using features based on manual annotations. For example, the performance of IMS HotCoref DE drops from 63.61

[5]Rosenberg et al. (2012) report good cross-language results of pitch accent detection on this dataset.

[6]The per-class accuracy is 82.1% for pitch accents and 37.1% for phrase boundaries. Despite these low quality phrase boundary annotations, we believe that, as a first step, their effectiveness can still be tested. This issue will be addressed in future work.

[7]www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/HOTCorefDe.html

to 48.61 CoNLL score[8] on the reference dataset TüBA-9 D/Z. The system, pre-trained on TüBA, yields a CoNLL score of 37.04 on DIRNDL with predicted annotations. This comparatively low score also confirms the assumption that the performance of a system trained on written text drops when applied to spoken text. The drop in performance can also be explained by the slightly different domains (newspaper text and radio news). However, if we train on the concatenation of the train and development set of DIRNDL we achieve a score of 46.11. This will serve as a baseline in the following experiments.

6 Experiments

We test our prosodic features by adding them to the feature set used in the baseline. We define *short NPs* to be of length 3 or shorter[9]. In this setup, we apply the feature only to short NPs. In the *all NP* setting, the feature is used for all NPs. The ratio of short vs. longer NPs in DIRNDL is roughly 3:1. Note that we evaluate on the whole test set in both cases. We report how the performance of the coreference resolver is affected in three settings:

(a) trained and tested on manual prosodic labels (short *gold*),

(b) trained on manual prosodic labels, but tested on automatic labels (this simulates an application scenario where a pre-trained model is applied to new texts (short *gold/auto*) and

(c) trained and tested on automatic prosodic labels (short *auto*).

Table 1 shows the effect of the pitch accent presence feature on our data. All features perform significantly better than the baseline[10]. As expected, the numbers are higher if we limit this feature to short NPs. We believe that this is due to the fact that the feature contributes most when it is most meaningful: on short NPs, a pitch accent makes it more likely for the NP to contain new information, whereas long NPs almost always have at least one pitch accent, regardless of its information status. We achieve the highest performance with gold labels, followed by the *gold/auto* version with a score that is not significantly worse than the *gold* version. This is important for applications as it suggests that the loss in performance is small when training on gold data and testing on predicted data. As expected, the version that is trained and tested on predicted data performs worse, but is still significantly better than the baseline. Hence, prosodic information is helpful in all three settings. It also shows that the assumption on short NPs in the pilot study is also true for automatic labels.

Baseline	46.11	
+ Accent	short NPs	all NPs
+ Presence gold	53.99	49.68
+ Presence gold/auto	52.63	50.08
+ Presence auto	49.13	49.01

Table 1: Pitch accent presence

Baseline	46.11	
+ Nuclear accent	short NPs	all NPs
+ Presence gold	48.63	52.12
+ Presence gold/auto	48.46	51.45
+ Presence auto	48.01	50.64

Table 2: Nuclear accent presence

Table 2 shows the effect of adding nuclear accent presence as a feature to the baseline. Again, we report results that are all significantly better than the baseline. The improvement is largest when we apply the feature to all NPs, i.e. also including long, complex NPs. This is in line with the findings in the pilot study for long NPs. If we restrict ourselves to just nuclear accents, this feature will receive the value *true* for only a few of the short NPs that would otherwise have been assigned *true* in terms of general pitch accent presence. Therefore, nuclear pitch accents do not provide sufficient information for a majority of the short NPs. For long NPs, however, the presence of a nuclear accent is more meaningful.

The performance of the different systems follows the pattern present for pitch accent type: *gold* $>$ *gold/auto* $>$ *auto*. Again, automatic prosodic information contributes to the system's performance. The highest score when using automatic labels is 50.64, as compared to 53.99 with gold labels. To the best of our knowledge, these are the best results reported on the DIRNDL anaphora dataset so far.

[8] We report the performance of the coreference system in terms of the CoNLL score, the standard measure to assess the quality of coreference resolution.

[9] In our experiments, this performed even better than length 4 or shorter as used in Rösiger and Riester (2015).

[10] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.01 level.

EXPERTEN	{der Großen KOALITION}$_1$	haben	sich	auf [...]	ein	Niedriglohn-
Experts	*(of) the grand coalition*	*have*	*themselves*	*on*	*a*	*low wage*

Konzept	VERSTÄNDIGT.	Die strittigen Themen [...]	sollten	bei	der	nächsten
concept	*agreed.*	*The controversial topics*	*shall*	*at*	*the*	*next*

Spitzenrunde	{**der Koalition**}$_1$	ANGESPROCHEN	werden.
meeting	*(of) the coalition*	*raised*	*be.*

EN: Experts within the the grand coalition have agreed on a strategy to address [problems associated with] low income. At the next meeting, ***the coalition*** *will talk about the controversial issues.*

Figure 2: Example from the DIRNDL dataset with English translation. The candidate NP (anaphor) of the coreference chain in question is marked in boldface, the antecedent is underlined. Pitch accented words are capitalised.

7 Analysis

In the following section, we discuss two examples from the DIRNDL dataset that provide some insight as to how the prosodic features helped coreference resolution in our experiments.
The first example is shown in Figure 2. The coreference chain marked in this example was not predicted by the baseline version. With prosodic information, however, the fact that the NP *"der Koalition"* is deaccented helped the resolver to recognise that this was given information: it refers to the recently introduced antecedent *"der Großen Koalition"*. This effect clearly supports our assumption that the absence of pitch accents helps for short NPs.

An additional effect of adding prosodic information that we observed concerns the length of antecedents determined by the resolver. In several cases, e.g. in Example (3), the baseline system incorrectly chose an embedded NP (1A) as the antecedent for a pronoun. The system with access to prosodic information correctly chose the longer NP (1B)[11]. Our analysis confirms that this is due to the accent on the short NP (on *Phelps*). The presence or absence of a pitch accent on the adjunct NP (on *USA*) does not appear to have an impact.

(3) {{Michael PHELPS}$_{1A}$ aus den USA}$_{1B}$. {**Er**}$_1$ ...
Michael Phelps from the USA. He ...

[11]The TüBA-D/Z guidelines state that the maximal extension of the NP should be chosen as the markable. http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-coreference-manual-2007.pdf

8 Conclusion and future work

We show that using prosodic labels that have been obtained automatically significantly improves the performance of a coreference resolver. In this work, we predict these labels using a CNN model and use these as additional features in IMS HotCoref DE, a coreference resolution system for German. Despite the quality of the predicted labels being slightly lower than the gold labels, we are still able to replicate results observed when using manually annotated prosodic information. This encouraging result also confirms that not only is prosodic information helpful to coreference resolution, but that it also has a positive effect even when predicted by a system.

A brief analysis of the resolver's output illustrates the effect of deaccentuation. Further work is necessary to investigate the impact on the length of the predicted antecedent.

One possibility to increase the quality of the predicted prosody labels would be to include the available lexico-syntactic information into the prosodic event detection model, since this has been shown to improve prosodic event recognition (Sun, 2002; Ananthakrishnan and Narayanan, 2008). To pursue coreference resolution directly on speech, a future step would be to perform all necessary annotations on automatic speech recognition output. As a first step, our results on German spoken text are promising and we expect them to be generalisable to other languages with similar prosody.

Acknowledgements

We would like to thank Kerstin Eckart for her help with the preparation of DIRNDL data. This work

was funded by the German Science Foundation (DFG), Sonderforschungsbereich 732, Project A6 and A8, at the University of Stuttgart.

References

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. 2012. Coreference in spoken vs. written texts: a corpus-based analysis. In *Proceedings of LREC*.

Sankaranarayanan Ananthakrishnan and Shrikanth S. Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical and syntactic evidence. In *IEEE Transactions on Audio, Speech and Language Processing*. volume 16, pages 216–228.

Stefan Baumann and Arndt Riester. 2013. Coreference, lexical givenness and prosody in German. *Lingua* 136:16–37.

Stefan Baumann and Anna Roth. 2014. Prominence and coreference – On the perceptual relevance of F0 movement, duration and intensity. In *Proceedings of Speech Prosody*. pages 227–231.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of LREC*. pages 3222–3228.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In Sebastian Nordhoff Christian Chiarcos and Sebastian Hellmann, editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, Springer, pages 65–76.

Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller. 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. pages 835–838.

Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. Technical Report ECS-95-001, Boston University.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*. Association for Computational Linguistics, pages 1–40.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Stroudsburg, PA, USA, SemEval '10, pages 1–8.

Andrew Rosenberg, Erica Cooper, Rivka Levitan, and Julia Hirschberg. 2012. Cross-language prominence detection. In *Speech Prosody*.

Ina Rösiger and Jonas Kuhn. 2016. IMS HotCoref DE: a data-driven co-reference resolver for German. In *Proceedings of LREC 2016*.

Ina Rösiger and Arndt Riester. 2015. Using prosodic annotations to improve coreference resolution of spoken text. In *Proceedings of ACL-IJCNLP*. pages 83–88.

Sidney Siegel and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.

Sabrina Stehwien and Ngoc Thang Vu. 2017. Prosodic event detection using convolutional neural networks with context information. In *Proceedings of Interspeech*.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. pages 168–175.

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Proceedings of ICSLP-2002*. pages 16–20.

Jacques Terken and Julia Hirschberg. 1994. Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech* 37(2):125–145.

Joel Tetreault and James Allen. 2004. Dialogue structure and pronoun resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium*.

Carla Umbach. 2002. (De)accenting definite descriptions. *Theoretical Linguistics* 2/3:251–280.