Coordination Boundary Identification without Labeled Data for Compound Terms Disambiguation

Yuya Sawada ♠ **Takashi Wada**♣ **Takayoshi Shibahara**♠ **Hiroki Teranishi**♠
Shuhei Kondo♠ **Hiroyuki Shindo**♠ **Taro Watanabe**♠ **Yuji Matsumoto**♢
♠ Nara Institute of Science and Technology
♣ School of Computing and Information Systems, The University of Melbourne
♢ RIKEN Center for Advanced Intelligence Project (AIP)
{yuya.sawada.sr7,shibahara.takayoshi.sk4,teranishi.hiroki.sw5,
shuhei-k,shindo,taro}@is.naist.jp
twada@student.unimelb.edu.au
yuji.matsumoto@riken.jp

Abstract

We propose a simple method for nominal coordination boundary identification. As the main strength of our method, it can identify the coordination boundaries without training on labeled data, and can be applied even if coordination structure annotations are not available. Our system employs pre-trained word embeddings to measure the similarities of words and detects the span of coordination, assuming that conjuncts share syntactic and semantic similarities. We demonstrate that our method yields good results in identifying coordinated noun phrases in the GENIA corpus and is comparable to a recent supervised method for the case when the coordinator conjoins simple noun phrases.

1 Introduction

In the scientific literature, coordination is a common syntactic structure and is frequently used to describe technical terminologies. These coordinate structures often involve ellipsis, a linguistic phenomenon in which certain redundant words inferable from the context are omitted. For instance, the phrase "*prostate cancer and breast cancer cells*" conjoins two cell names, "*prostate cancer cell*" and "*breast cancer cell*," with the token "cell" eliminated from the first conjunct. This phenomenon raises significant challenges in named entity recognition (NER) tasks, and most of the current NER models (Ma and Hovy, 2016) can identify only non-elliptical conjuncts, e.g., "*breast cancer cells*," or incorrectly extract the whole coordinate phrases as single complex entities. Therefore, identifying coordinated noun phrases is crucial to improving the model performance in NLP, particularly in NER and relation extraction tasks within in scientific domains.

In this paper, we propose a simple yet effective method for finding coordination with related compound nouns, such as technical terms. Compared to previous methods (Ficler and Goldberg, 2016b; Teranishi et al., 2017; Teranishi et al., 2019), our approach does not require any training on labeled data, and is applicable under the realistic conditions where annotations of coordinate structures are not readily available. Our method employs recent pre-training language models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) to measure the similarities of words, and identifies coordination boundaries based on the property in which conjuncts share syntactic and semantic similarities. This property has been exploited in traditional alignment-based methods (Kurohashi and Nagao, 1994; Shimbo and Hara, 2007; Hara et al., 2009), and our system extends and simplifies such methods by using neural embedding representations instead of the handcrafted features or heuristic rules used in their approaches. Our experiments show that, even without training, our method achieves good results in identifying nominal coordination boundaries in the GENIA corpus (Tateisi et al., 2005). When targeting only the coordination of noun phrases that do not contain clauses or prepositional phrases, our method is even comparable to a supervised baseline model trained on annotated data.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3043–3049
Barcelona, Spain (Online), December 8-13, 2020

2 Related Studies

The goal of our method is to identify the coordination boundaries of noun phrases, including scientific named entities. In this respect, our work is similar in spirit to studies by Buyko et al. (2007) and Chae et al. (2014). Buyko et al. (2007) proposed a supervised method trained on coordination-annotated data and resolved coordination ellipses included in biomedical named entities. In addition, Chae et al. (2014) proposed a dictionary-based method that resolves complex ellipses in coordinated noun phrases using linguistic rules and an entity mention dictionary. However, such annotated data or dictionaries are not readily available in practice. Therefore, we propose a new method that does not require any training on the labeled data, and is applicable under realistic scenarios.

In terms of the methodology, our study is highly inspired by the traditional alignment-based approaches for identifying the scopes of coordinate structures (Kurohashi and Nagao, 1994; Shimbo and Hara, 2007; Hara et al., 2009). Their methods identify coordination boundaries by aligning similar words or phrasal units before and after a coordinator, assuming that the conjuncts share semantic and syntactic similarities. Our method extends and simplifies their models by replacing their handcrafted features and heuristic rules with recent word embeddings. Our method also differs from the studies by Shimbo and Hara (2007) and Hara et al. (2009) trained the weights on handcrafted features using annotated data, whereas our method employs pre-trained word embeddings and does not require any training on the labeled data.

3 Proposed Method

Algorithm 1: Our System Flow

Input: $\{w_1, w_2, ..., w_N\}$ and k
Output: best_span
i, j = Preprocess($w_{1:N}, k$);
best_score = $-\infty$;
best_span = ϕ;
for b = *i to k - 1* **do**
  score, span = Alignment($w_{b:k-1}, w_{k+1:j}$);
  if *score* $>$ *best_score* **then**
    best_score = score;
    best_span = span;
  end
end

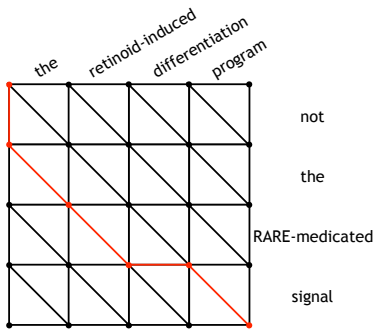

Figure 1: An example of an edit graph

Given a sentence that consists of N words $w_{1:N} = \{w_1, w_2, ..., w_N\}$ and the coordinator w_k, the goal of our method is to identify the span of the two noun conjuncts adjoined by w_k. We presume that the first conjunct ends at w_{k-1} and the second conjunct starts at w_{k+1}. Our model predicts a span $(i, j | i < k < j)$, where w_i is the beginning word of the first conjunct, and w_j is the end word of the second conjunct. To achieve this, our method involves two procedures: preprocessing and sequence alignment. The pseudo-code is shown in Algorithm 1. Our method first suggests the candidates of noun conjuncts using a few simple rules and applies a sequence alignment to determine the best span. In the following, we describe each procedure in detail.

3.1 Preprocessing

During preprocessing, we delimit the possible spans of the conjuncts from a coordinator word (e.g., "and," "or," "but"). We extract the longest spans before and after a coordinator that does not contain the following types of words or tokens: a verb, preposition, or certain punctuation marks, i.e., a comma, colon, semicolon or ellipsis (...). For instance, given a sentence "*We show that induction of a trimer of the NFAT and Oct sites is not sensitive to phorbol ester treatment*," the sequences "*the NFAT*" and

"*Oct sites*" are retrieved as the longest candidate spans of the conjuncts. Using the sequence alignment technique described below, our method aims to identify the correct conjunct pair, "*NFAT and Oct*".

3.2 Sequence Alignment

Following Shimbo and Hara (2007), we employ the sequence alignment technique (Levenshtein, 1966) to determine the best span of the coordination. A sequence alignment is a method that transforms one sequence into another through a series of edit operations, namely "deletion," "insertion," and "substitution." In this study, we define both a deletion and insertion as "skipping," and a substitution as an "alignment." We calculate the scores of alignments based on the similarities of words and assign a constant value to the skipping operation. The optimal alignment with the maximum score is computed by dynamic programming in a lattice graph, called an edit graph. Figure 1 illustrates an example of the edit graph for the conjunct candidate, "*the retinoid-induced differentiation program but not the RARE-mediated signal*." A diagonal edge represents the alignment between two words at the top and right of the edge. In this example, three pairs of words ("the–the," "retinoid-induced–RARE-mediated," and "program–signal") are aligned. The vertical and horizontal edges represent a skipping operation, which indicates that the words are not aligned. To calculate the similarities of the words, we use the square of the cosine similarity of the word embeddings.[1] We used three different word embedding methods, namely, BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), and FastText[2] (Bojanowski et al., 2017) to investigate the impacts of the embedding methods. For BERT and ELMo, we input a whole sentence containing a conjunction, and use the last layer of the hidden states as the contextualized word embeddings.[3]

The largest difference between the approach by Shimbo and Hara (2007) and our method is that they use coordination-annotated data to train the feature weights, whereas our method does not. This difference requires some modifications in their algorithm: Because we do not have access to the gold span of the conjuncts, we need to consider all possible candidates of conjuncts within the outer boundaries, which are determined using our preprocessing step. However, because it is computationally expensive to consider all combinations of the conjuncts, we always fix the row of the edit graph as the longest span of the second conjunct, and change the column with different candidates for the first conjunct.[4] For instance, for the extracted outer boundary "*the NFAT and Oct sites*," we always set "*Oct sites*" as the row of the edit graph, and create two graphs whose columns are "*NFAT*" and "*the NFAT*," respectively. We then select the path with the best score for each graph, resulting in two optimal paths and scores in this example. To predict the best span, we choose the best path with the highest score among the multiple paths obtained from each edit graph. To take into account the differences in path lengths, we normalize the score by the path length to the power of a constant value of between 0.0 and 1.0, which we tuned using the development data.[5,6]

4 Experiment

4.1 Baseline

We compared our method with the latest strong supervised model (Teranishi et al., 2019). Their method trains bidirectional long short-term memories (BiLSTMs) (Hochreiter and Schmidhuber, 1997) on annotated data, and runs the CKY algorithm to find the globally optimal coordinate structures in a sentence.

4.2 Dataset

We evaluate our method on GENIA treebank beta (Tateisi et al., 2005), which is a biomedical-domain corpus that consists of abstracts taken from the MEDLINE database, and contains syntactic annotations,

[1] When the similarity takes a negative value, we multiply the square by -1.

[2] We used word embeddings pre-trained in bio-domain corpora, namely SciBERT (Beltagy et al., 2019) and Biowordvec (Yijia et al., 2019). For ELMo, we used the model trained on PubMed, available at `https://allennlp.org/elmo`.

[3] When there is a word decomposed into subwords, we create the word vector from the mean of the subword vectors.

[4] The span of the second conjunct is determined by the path of the edit graph; once the path reaches the right-most column, we stop the operation and regard the last vertex as the span of the second conjunct.

[5] We tune our hyper-parameters, namely, the skip score and normalization value, on the extended Penn Treebank (Ficler and Goldberg, 2016a).

[6] Note that the span for the first conjunct should be enumerated owing to this length normalization.

including coordination phrases. In total, it contains 2508 sentences, and a nominal coordination is included in the 2317 sentences. Following Teranishi et al. (2019), we apply a 5-fold cross-validation and take the average performance on the held-out data over five runs. Unlike the supervised baseline, however, our method does not require any training, and therefore does not use any held-in data[7]. During the inference, gold POS tags are used for both the baseline and our model; our model uses them during our preprocessing step, as described in Section 3.1.

4.3 Evaluation

We compare the baseline and our method based on the recall of the predicted spans of a nominal coordination, as in Teranishi et al. (2019). That is, we evaluate how well the models can identify the correct spans. We calculate the scores in two cases i.e., when we target all nominal coordination structures (All NP), and when we only deal with simple nominal coordination structures (Simple NP) that do not contain any prepositional phrases, clauses, or special punctuation marks that we use to delimit the span of conjuncts in our preprocessing.

4.4 Results

	All NP	Simple NP
Ours (Biowordvec)	0.447	0.556
Ours (ELMo)	0.602	0.748
Ours (SciBERT)	0.561	0.697
Teranishi+19 (paper)	**0.706**	-
Teranishi+19 (code)	0.695	**0.766**

Table 1: Recall with GENIA.

	Simple NP		
	P	R	F
default (ELMo)	0.607	**0.748**	0.670
+threshold	0.624	0.728	0.672
+rules	0.612	**0.748**	0.673
+rules+threshold	**0.705**	0.719	**0.712**

Table 2: Precision, recall, and F1 scores of identifying coordination boundaries of simple noun phrases on GENIA with additional rules.

Table 1 shows the results of the baseline and our proposed method. As the table indicates, our method can identify the coordination boundary with a good level of accuracy, given that it does not conduct any training on the labeled data. When we focus on a Simple NP, our model is even comparable to the strong supervised baseline. Focusing on the results of the three different word embeddings, the contextualized word embeddings significantly outperform the static ones in terms of word alignment. ELMo performs better than SciBERT, and we conjecture that this would be because ELMo employs character-level CNN to encode words and works well for aligning biological terminologies that have similar character strings. Similarly, Claudia and Damir (2020) have shown that ELMo which is trained on PubMed produces better word embeddings than SciBERT for measuring similarities of medical terms.

To consider more realistic conditions, we conducted another experiment in which the system has to identify whether the coordinated phrases are noun phrases or not. During this experiment, we added a few heuristic rules to our model to enhance its ability to identify the type of coordination. First, we simply discard the cases when adjectives or adverbs are adjutant to the coordinator. We also discard the conjuncts whose scores are below a certain threshold; this rule helps eliminate the cases when a conjunction conjoins two sentences, and not phrases, for instance. Table 2 shows the precision, recall, and F1 measures for a simple NP.[8] It shows that, with a few heuristics, our method can detect the nominal coordination and identify the boundaries with good recall and precision. This result suggests that our model can potentially be applied to more realistic and challenging problems, such as the resolution of ellipses in scientific literature.

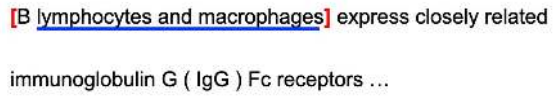

Figure 2: A representative error of our model using ELMo. The underlined span indicates the model prediction, and the bracketed span represents the gold annotation.

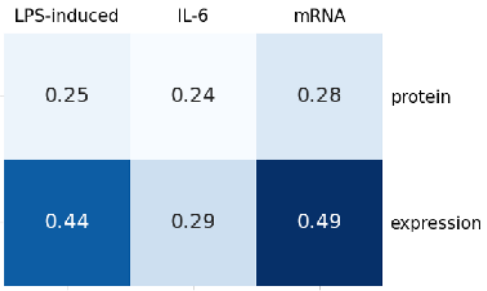

Figure 3: An example of the similarity table generated by our method using SciBERT. Columns and rows indicate the left and right conjunct candidates, respectively.

4.5 Error Analysis

To better understand the behavior of our model, we perform an error analysis. Figure 2 shows an example when our model using ELMo fails to identify the coordination boundary: the correct span is "*B lymphocytes and macrophages*" whereas our model mistakenly identifies the span as "*lymphocytes and macrophages*." What makes it very challenging to solve this case is the inequality of the number of words before and after the coordinator; because the sequence alignment algorithm we use assumes one-to-one alignment, our model tends to struggle finding one-to-many alignment such as "B lymphocytes"–"macrophages".

When we compare our model with different word embedding methods, SciBERT sometime produces close representations for unrelated words and leads to erroneous alignment. Figure 3 shows an example of such a case with a similarity table. In this case, our model using ELMo correctly predicts the span as "*mRNA and protein*," although when applied to SciBERT it extracts the longer span "*IL-6 mRNA and protein expression*." This error stems from the inaccuracy of the similarity table, where "*mRNA*" and *"expression"* is more similar than the correct alignment "*mRNA*" and *"protein"*.

5 Conclusion

In this study, we proposed a simple yet effective method for identifying the coordination boundary of noun phrases. Our method identifies the coordination boundary by aligning words before and after a coordinator, assuming that they should share syntactic and semantic similarities. To calculate word similarities, our method exploits recent word embedding methods and finds the optimal alignment using the sequence alignment technique. Our experiments on the GENIA corpus show that, without any training, our method can identify the coordination boundaries of noun phrases with good accuracy. When looking at the coordination that conjoins simple noun phrases, our method is comparable to the strong supervised model trained on annotated data. We also show that, using a few heuristic rules, our model can identify noun conjuncts with good recall and precision. This result suggests that our model can potentially be applied to more realistic problems, such as the resolution of ellipses in the scientific literature. In a future study, we plan to improve the performance of identifying such structures and integrate our method into the NER system.

[7]We did not use the coordination-annotated Penn Treebank (Ficler and Goldberg, 2016a) for the experiments due to the small number of elliptical coordination structures found in the data: its development data contain only 439 nominal coordination structures among a total of 848, and the number of elliptical coordination structures is as few as 74. There are also many terms that are ambiguous in term of whether they are a compound or not, such as "*House Ways and Means Committee*".

[8]Precision indicates how well the model identifies noun phrases *and* predicts their correct spans.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ekaterina Buyko, Katrin Tomanek, and Udo Hahn. 2007. Resolution of coordination ellipses in biological named entities using conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 163–171.

Jeongmin Chae, YoungHee Jung, Taemin Lee, Soonyoung Jung, Chan Huh, Gilhan Kim, Hyeoncheol Kim, and Heung-Bum Oh. 2014. Identifying non-elliptical entity mentions in a coordinated np with ellipses. *Journal of biomedical informatics*, 47:139–152, February.

Schulz Claudia and Juric Damir. 2020. Can embeddings adequately represent medical terminology? new large-scale medical term similarity datasets have the answer! In *Proceeding of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8775–8782, New York, USA, February. Association for the Advancement of Artificial Intelligence.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2016a. Coordination annotation extension in the Penn Tree Bank. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 834–842, Berlin, Germany, August. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2016b. A neural network for coordination boundary prediction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 23–32, Austin, Texas, November. Association for Computational Linguistics.

Kazuo Hara, Masashi Shimbo, Hideharu Okuma, and Yuji Matsumoto. 2009. Coordinate structure analysis with global structural constraints and alignment-based local features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 967–975, Suntec, Singapore, August. Association for Computational Linguistics.

Sepp Hochreiter and Júrgen' Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 10:707–710.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Masashi Shimbo and Kazuo Hara. 2007. A discriminative learning model for coordinate conjunctions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 610–619, Prague, Czech Republic, June. Association for Computational Linguistics.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Coordination boundary identification with similarity and replaceability. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 264–272, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Decomposed local models for coordinate structure parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3394–3403, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhang Yijia, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6, December.