Improving a Strong Neural Parser with Conjunction-Specific Features

Jessica Ficler
Computer Science Department
Bar-Ilan University
Israel
jessica.ficler@gmail.com

Yoav Goldberg
Computer Science Department
Bar-Ilan University
Israel
yoav.goldberg@gmail.com

Abstract

While dependency parsers reach very high overall accuracy, some dependency relations are much harder than others. In particular, dependency parsers perform poorly in coordination construction (i.e., correctly attaching the *conj* relation). We extend a state-of-the-art dependency parser with conjunction-specific features, focusing on the similarity between the conjuncts head words. Training the extended parser yields an improvement in *conj* attachment as well as in overall dependency parsing accuracy on the Stanford dependency conversion of the Penn TreeBank.

1 Introduction

Advances in dependency parsing result in impressive overall parsing accuracy. For the most part, the advances are due to general improvements in parsing technology or feature representation, and do not explicitly target any specific language or syntactic construction. However, despite the high overall accuracy, parsers are still persistently wrong in attaching certain relations. In the attachments predicted by BIST-parser (Kiperwasser and Goldberg, 2016), the F1 score for the labels *nn*, *nsubj*, *pobj*, and others is 95% and above; while the F1 scores for *advmod*, *conj* and *prep* are 83.3%, 82.5% and 87.4% respectively. Conjunction holds the lowest F1 score, ignoring rare labels, *dep* and *punct*. Other parsers behave similarly. Conjunction mistakes occurs also in simple sentences such as:

(1) "Those machines are still considered novelties, with keyboards only a munchkin could love and screens to match."

(2) "In the year-earlier period, CityFed had net income of $ 485,000, but no per-share earnings."

BIST-parser (Kiperwasser and Goldberg, 2016) attaches *screens* and *love* instead *screens* and *keyboards* in (1); and *earnings* and *had* instead *earnings* and *income* in (2).

The parsers low performance on conjunction is disappointing given that conjunction is a common and important syntactic phenomena, appearing in almost 40% of the sentences in the Penn TreeBank (Marcus et al., 1993), as well constitutes 2.82% of the Stanford dependency conversion of the Penn TreeBank (De Marneffe and Manning, 2008) edges.

In this work we focus on improving *conj* attachment accuracy by extending a dependency parser with features that specifically target the coordinating conjunction structures. Similar efforts were done for constituency parsing in previous work (Hogan, 2007; Charniak and Johnson, 2005).

As previously explored, conjuncts tend to be semantically related and have a similar syntactic structure (Shimbo and Hara, 2007; Hara et al., 2009; Hogan, 2007; Ficler and Goldberg, 2016; Charniak and Johnson, 2005; Johnson et al., 1999). For example: *"for China and for India"*, *"1.86 marks and 139.75 yen"*, *"owns 33 % of Moleculons stocks and holds 27.5 % of Datapoints shares"*. Such cases are common but still there are many cases where symmetry between conjuncts is less straightforward such as in (1), which includes the conjuncts *"keyboards only a munchkin could love"* and *"screens to match"*; and (2), which includes *"net income of $ 485,000"* and *"no per-share earnings"*. For many cases of this type, the head words of the conjuncts are similar, e.g. *(keyboards,screens)* in (1) and *(income,earnings)* in (2).

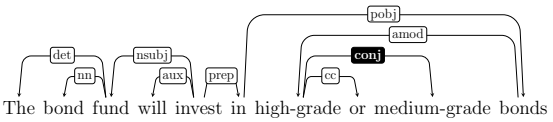
We extend BIST-parser, the Bi-LSTM based

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 343–348,
Valencia, Spain, April 3-7, 2017.

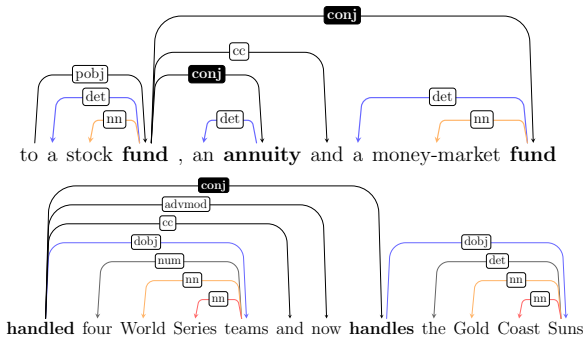
parser by Kipwasser and Goldberg (2016), by adding explicit features that target the conjunction relation and focus on various aspects of symmetry between the potential conjuncts' head words. We show improvement in dependency parsing scores and in *conj* attachment.

2 Symmetry between Conjuncts

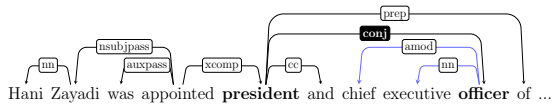
It is well known that conjuncts tend to be semantically related and often have a similar syntactic structure. This property of coordination was used as a guiding principle in previous work on coordination disambiguation (Hara et al., 2009; Hogan, 2007; Shimbo and Hara, 2007; Ficler and Goldberg, 2016). While these focus on symmetry between conjuncts in constituency structures, we use the symmetry assumption for the purpose of improving dependency parsing. Here is a simple example of dependency tree that include conjunction:



The edge labeled with *conj* connects the first conjunct head to the heads of the other conjuncts. In more complex conjuncts, the subtrees under the nodes connected by *conj* are often similar such as the following examples:



However, there are also cases where the conjuncts structures are non-similar such as in:



Yet, some form of symmetry (or anti-symmetry) usually holds between the conjuncts head words. Table 1 lists the most common coordinated words in the PTB.

	(Head,Modifier)		
1.	($,$)	13.	(chairman,executive)
2.	($,cents)	14.	(on,on)
3.	(president,officer)	15.	(by,by)
4.	(%,%)	16.	(at,at)
5.	(chairman,officer)	17.	($,%)
6.	(securities,exchange)	18.	(marks,yen)
7.	(in,in)	19.	(president,executive)
8.	(standard,poor)	20.	(savings,association)
9.	(to,to)	21.	(chairman,president)
10.	(buy,sell)	22.	(inc.,inc.)
11.	(for,for)	23.	(from,from)
12.	(corp.,corp.)	24.	(shares,%)

Table 1: The most common *conj* attachments in the Penn TreeBank dependency conversion.

3 Conjunction Features

We suggest a set of features that are designed specifically for the conjunction relation, and target the symmetry aspect of the head words. The features look at a pair of head and modifier words, and are based on properties that appear frequently in conjunctions in the Stanford Dependencies version of the PTB. The features are summarized in Table 2, and are detailed below:

CAP – The case where both conjuncts head words start with a capital letter is much more common ($> 3\times$) than the case where only one of the head words starts with a capital letter. These cases are usually names of people, countries and organizations; and common phrases such as *"Mac and Cheese"*. This property is rare in other labels except *nn*. We capture this property with a boolean feature that indicates whether both conjuncts head words start with a capital letter.

SUF – In some of the conjunctions, the head words have a similar form, as in (codification, clarification), (demographic, geographic), (high-grade, medium-grade), (backwards, forwards). The cases where the longest common suffix between the words is at least 3 is 8% in the case of *conj* and much lower for the other labels. We capture this tendency using a numeric feature that indicated the length of the common suffix between the head words.

LEM – Conjuncts heads often share the same lemma. These are usually different inflections of the same verb (e.g. sells,sold); or singular/plural forms of the same noun (e.g. table,tables). This is also a tendency that is more common in *conj* label than the other labels. We capture these, with a boolean feature indicating whether the lemmas of

	Description	Type	Examples
CAP	Whether both words start with a capital letter	boolean	(Corp.,Inc.), (Poland,Hungary)
SUF	The length of the longer common suffix between the words	numeric	(men,women), (three-month,six-month)
LEM	Whether the words lemmas are identical	boolean	(say,said), (handled,handles)
SYM	The cosine distance between the words embeddings	numeric	(reported,said), (president,director)
SENT-H	Whether the head word sentiment is positive, negative or neutral	1,-1, or 0	(up,down), (confirmed,declined)
SENT-M	Whether the modifier word sentiment is positive, negative or neutral	1,-1, or 0	

Table 2: Summary of the conjunction-specific features.

the conjuncts head words are identical. Lemmas are obtained using the NLTK (Bird, 2006) interface to WordNet (Miller, 1995).

SYM – The conjuncts head words usually have a strong semantic relation. For example (fund, annuity), (same, similar), (buy, sell), (dishes, glass). `SYM` is a numeric feature that scores the similarity between the conjuncts heads words. The score is computed as the cosine-similarity between word embeddings of the head words (these embeddings are initialized with pre-trained vector from Dyer et al. (2015)).

SENT – In some cases, both conjunct's head words sentiments are not neutral. Here are some examples from the PTB where both words are with positive sentiments: (enjoyable,easy), (complementary,interesting), (calm,rational); where both words are with negative sentiments: (slow,dump), (insulting,demeaning), (injury,death); and where one word is positive and the other is negative: (winners,losers), (crush,recover), (succeeded,failed). Having non-neutral sentiment for both words is not very common for *conj* relation (2.3% of the cases), but it much less common for the other relations. Therefore we add features that indicate the sentiment (positive, negative or neutral) for each of the coordinated words. We use lists of positive and negative words from work on airline consumer sentiment (Breen, 2012).

4 Incorporating conjunction features

We incorporate the above features in the freely available BIST-parser (Kiperwasser and Goldberg, 2016). This parser is a greedy transition-based parser, using the archybrid transition system (Kuhlmann et al., 2011). At each step of the parsing process, the parser chooses one of $2*|labels|+1$ possible transitions: SHIFT, $\text{RIGHT}_{(\text{rel})}$ and $\text{LEFT}_{(\text{rel})}$. The LEFT and RIGHT transitions add a dependency edge with the label `rel`. At each step, all transitions are scored, and the highest scoring transition is applied.

The Stanford Dependencies scheme specifies that the *conj* relation appears as a right edge, and so it can only be produced by a $\text{RIGHT}_{(\text{conj})}$ transition. We compute a score S_{conj} which is added to the score of the $\text{RIGHT}_{(\text{conj})}$ transition that was produced by the parser. S_{conj} is computed by an MLP that receives a feature vector that is a concatenation of the original parser's features and the conjunction specific features. The scoring MLP and the parser are trained jointly.

5 Experiments

We evaluate the extended parsing model on the Stanford Dependencies (De Marneffe and Manning, 2008) version of the Penn Treebank. We adapt BIST-parser code to run with the DyNet toolkit[1] and add our changes. We follow the setup of Kiperwasser and Goldberg (2016): (1) A word is represented as the concatenation of randomly initialized vector and pre-trained vector (taken from Dyer et al. (2015)); (2) The word and POS embeddings are tuned during training; (3) Punctuation symbols are not considered in the evaluation; (4) The hyper-parameters values are as in Kiperwasser and Goldberg paper (2016), Table 2; (5) We use the same seed and do not perform hyper-parameter tuning. We train the parser with the conjunction features for up to 10 iterations, and choose the best model according to the LAS accuracy on the development set.

General Parsing Results Table 3 compares our results to the unmodified BIST parser. The extended parser achieves 0.1 points improvement in UAS and 0.2 points in LAS comparing to Kiperwasser and Goldberg (2016). This is a strong baseline, which so far held the highest results among greedy transition based parsers that were trained on the PTB only, including e.g. the parsers of Weiss et al (2015), Dyer et al (2015) and Ballesteros et al (2016). Stronger absolute parsing numbers are reported by Andor et al (2016) (using a beam); and Kuncoro et al (2016) and Dozat

[1] `https://github.com/clab/dynet`

and Manning (2016) (using an arc-factored global parsers). All those parsers rely on broadly the same kind of features, and while we did not test this, it is likely the conjunction features would benefit them as well.[2]

Parsing Results for *conj* Label We evaluate our model specifically for *conj* label, and compare to the results achieved by the parser without the conjunction features. We measure Rel (correctly identifying modifiers that participate in a conj relation, regardless of correctly attaching the parent) and Rel+Att (correctly identifying both the head and the modifier in a conj relation). The results are in Table 4. The improvement in Rel score is relatively small while there is an improvement of 1.1 points in Rel+Att F1 score, suggesting that the parser was already effective at identifying the modifiers in a conj relation and that our model's benefit is mainly on attaching the correct parent node.

Analysis We would like to examine to what extent the improvement we achieve over Kiperwasser and Goldberg (2016) on *conj* attachments corresponds to the coordination features we designed. To do that, we analyze the *conj* cases in the dev-set that were correctly predicted by our model and were not predicted by the original BIST-parser and vice versa. The following table shows the percentage of cases where conjunction features appear in each of these lists:

Features	+Our,−K&G	−Our,+K&G
LEM+CAP+SUF	7.5	0
LEM+SUF	3	0
SENTIMENT+SUF	1.5	0
LEM/CAP/SENTIMENT/SUF	29.9	24
Total	41.9	24

The percentage of cases that include conjunction features is much higher in the list of cases that were correctly predicted only by our model. More than that, there are no cases that include more than one conjunction feature in the list of cases that were correctly predicted only by BIST-parser (Kiperwasser and Goldberg, 2016).

[2]A reviewer of this work suggested that our baseline model is oblivious to the word's morphology, and that a neural parsing architecture that explicitly models the words' morphology through character-based LSTMs, such as the model of (Ballesteros et al., 2015), could capture some of the information in our features automatically, and thus would be a better baseline. While we were skeptical, we tried this suggestion, and found that it indeed does not change the results in a meaningful way.

System	UAS	LAS
Kiperwasser16	93.9	91.9
Kiperwasser16 + conjunction features	**94**	**92.1**

Table 3: Parsing scores on the PTB test-set (Stanford Dependencies).

	Kiperwasser16	Kiperwasser16 + conjunction features
Rel R	92.5	**92.9**
Rel P	**91.6**	91.5
Rel F1	92	**92.2**
Rel+Att R	83	**84.2**
Rel+Att P	82.1	**83**
Rel+Att F1	82.5	**83.6**

Table 4: Test-set results for *conj* label only.

The above table does not include the SYM feature since unlike the other features there is no absolute way to determine whether the feature takes place on a specific example. To give a sense of the contribution of the SYM feature, we show some examples where our model attaches a *conj* label between similar words, while the unmodified BIST parser attaches *conj* parent which is clearly less similar to the modifier (The word in bold is the attached modifier; the word marked with continuous line is the node's parent in our prediction; the word marked with dashed line is the node's parent in BIST's prediction):

- Koop, who rattled liberals and **conservatives** alike with his outspoken views on ...
- ... dropped in response to gains in the stock market and **losses** in Treasury securities.
- Died: Cornel Wilde, 74 ,actor and **director** ,in Los Angeles ,of leukemia ...
- ... investment firms advising clients to boost their stock holdings and **reduce** the „.

In the cases that were correctly predicted by BIST-parser only, we could not find examples where the words in the correct attachment are clearly more similar than the attachment predicted by our model. We could find a few examples where both models attached words that are similar, such as:

- ML & Co.'s net income dropped 37%, while BS Cos. posted a 7.5% gain in net, and PG Inc.'s profit fell, but would have **risen** ...
- The closely watched rate on federal funds, or overnight **loans** between banks, slid to...

6 Conclusions

While most recent work in parsing attempt to improve results using "general" architectures and feature sets, targeted feature engineering is still beneficial. We demonstrate that a linguistically motivated and data-driven feature-set for a specific syntactic relation (coordinating conjunction) improves a strong baseline parser.

The features we propose explicitly model the symmetry between the head words in coordination constructions. While we demonstrated their effectiveness in a greedy transition-based parser, the information our features capture is not currently captured also by other dependency parsing architectures (including first-order graph based parsers, higher-order graph-based parsers, beam-based transition parsers). These features will be straightforward to integrate into such parsers, and we expect them to be effective for them as well.

Acknowledgments

This work was supported by The Israeli Science Foundation (grant number 1555/15) as well as the German Research Foundation via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proc. of ACL*.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal, September. Association for Computational Linguistics.

Miguel Ballesteros, Yoav Goldberg, Chris Dyer, and Noah A Smith. 2016. Training with exploration improves a greedy stack-LSTM parser. In *proceedings of Short Papers EMNLP*.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Jeffrey Oliver Breen. 2012. Mining twitter for airline consumer sentiment. *Practical text mining and statistical analysis for non-structured text data applications*, 133.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. of ACL*, pages 173–180, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. In *arXiv preprint arXiv:1611.01734*.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China, July. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2016. 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.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proceedings of Association for Computational Linguistics*, pages 680–687.

Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic "unification-based" grammars. In *Proc. of ACL*, pages 535–541, College Park, Maryland, USA, June. Association for Computational Linguistics.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Marco Kuhlmann, Carlos Gómez-Rodríguez, and Giorgio Satta. 2011. Dynamic programming algorithms for transition-based dependency parsers. In *Proc. of ACL*, pages 673–682. Association for Computational Linguistics.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, and Noah A. Smith. 2016. Distilling an ensemble of greedy dependency parsers into one MST parser. In *Proc. of EMNLP*, pages 1744–1753, Austin, Texas, November. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

George A. Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China, July. Association for Computational Linguistics.