Dependency Parsing of Low-Resource Languages

Jiaming Luo, Sunoo Park, Hayk Saribekyan

- State-of-the-art parsers on English reach an F1-score of above 90%, but it is **hard to reach such scores on less common languages**.
- **Problem:** Lack of availability of **gold treebanks** on which to train parsers.
- **Idea: “**Transfer” dependency trees between different languages.

Prior work on multilingual dependency parsing

- Either assume availability of **multilingual parallel corpora**, which is **not available for low-resource languages**; Or improve unsupervised method by modelling structural similarity across languages.

Model	Dataset	UAS on German*
Self-Training Parsers [Rasooli, Collins 2015]	Universal Dependency Treebank Europarl (as parallel corpus)	Self-Training with Parallel data: 78.8%
Discriminative Transfer Parsers [Täckström et al. 2013]	CoNLL-X 2006 Dataset	Sharing Based on Language Groups: 59.2%
		Target Adaptation: 62.5%
Hierarchical Tensor Model [Zhang, Barzilay 2015]	Universal Dependency Treebank	Unsupervised: 62.5%
		Semi-supervised (50 sentences): 74.2%

* Presenting UAS on German as we have used German as a target language in our work

Our approach in a nutshell

Develop methods to project dependency arcs onto target-language sentences from **comparable** (or “almost parallel”) corpora?

1. **Identify** similar sentence pairs across **multilingual Wikipedia articles**, and **align** the words in these sentence pairs.
2. **Project dependency trees** from resource-rich language sentences to similar sentence in a resource-poor target language.
3. **Train an (MST) dependency parser** based on the projected (sub)trees.

Corpora

- Comparable text corpus: **Wikipedia** (en, fr, de)
- Parallel text corpus: **Europarl** (21 European languages [Koehn 2005])
- Dictionary: **Wiktionary** (en, fr)

Step 1. **Aligning** comparable sentence pairs

We consider Wikipedia articles that exist in both source and target languages.

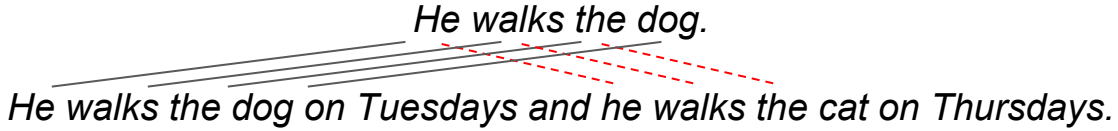
```
For each sentence in the source-language article:
    For each sentence in the target-language article:
        Align similar subtrees in the two sentences
        Project dependency arcs from source to target
```

Traditional alignment methods are not applicable because:

- They are designed for **parallel** sentences (i.e. exact translations).
- We want to take **any** two sentences and align the parts which are “similar”.

Choosing the **best match** in the target sentence

For any source word, there may be multiple matches in the target sentence.



Our method determines the best match for a source word by considering:

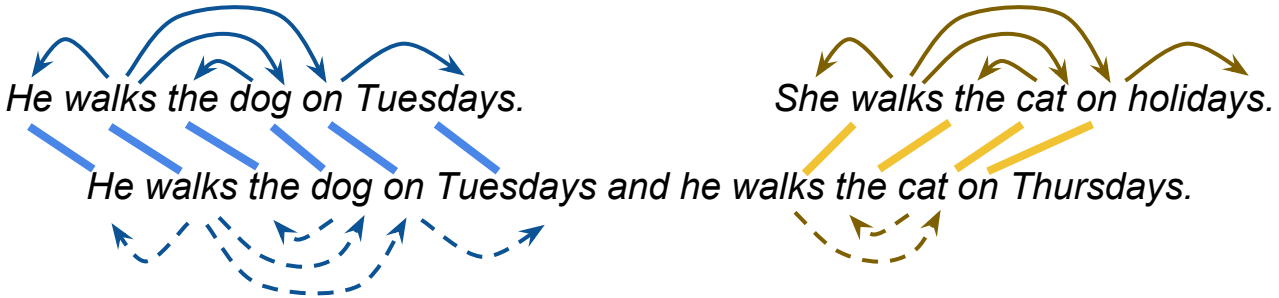
- Similarity of the **neighborhood of aligned words** in the source and target sentences (using *Manhattan* distance).
- Similarity of **contexts** represented as a **concatenation** of word vectors* appearing in the context window (using *cosine* distance).

*Since word vectors for different languages are trained using different monolingual corpora, we have to map both sets (source and target) of word vectors to the same semantic space [Faruqui, Dyer 2014].

Step 2. **Projecting** dependency trees

Given an **aligned, comparable sentence pair**, we project dependencies from the source-language sentence to the target-language sentence.

We may have **multiple projected (sub)trees**!



What if the projected arcs conflict?
(i.e. they don't form a tree/forest on the target sentence)

We decide which arcs to remove based on a **weight function** which can depend on the sizes of projected subtrees, and similarity of the source/target languages.

Results and **Analysis**

- Train an MST parser on the (partial) projections [McDonald et al. 2005].
- Two relevant metrics: **precision** of projections, and parsing **accuracy**.

Dataset	Precision*	Accuracy*
Europarl (50K)**	~75%	61%
Wikipedia (5K)***	~64%	41%

* The numbers presented are not final, since due to computation constraint, we did not run our model on the whole corpora.
** To put it into perspective, Europarl has about 2M parallel sentences. Note that our method does not make use of the fact that sentences are parallel; we only included Europarl here to see how much the result depends on the corpus quality.
*** We only scraped ~80K articles from Wikipedia, and trained the parser on 5K articles.

- With only a small subset of available comparable corpora, we rivalled the results from several SOTA unsupervised methods. We expect the numbers to further go up when leveraging a full-sized corpus.

Future directions

- Evaluate performance based on a **larger number of source languages**.
- Evaluate performance **when the target language is linguistically less close** to the source languages.
  - Important for resource-poor languages!
- Try to improve performance by experimenting with:
  - Dictionary quality
  - Algorithm to decide which projected trees to keep in case of conflicts
  - Improving projection methods to get more projected arcs (important for accuracy!)
  - Different types of parsers (e.g. RBG)

References

[Bird et. al 2009] **Natural Language Processing with Python**. *Steven Bird, Edward Loper and Ewan Klein*. O'Reilly Media Inc.
[Buchholz, Marsi 2006] **CoNLL-X shared task on Multilingual Dependency Parsing**. *Sabine Buchholz and Erwin Marsi.* CoNLL-X 2006.
[Faruqui, Dyer 2014] **Improving Vector Space Word Representations Using Multilingual Correlation**. *Manaal Faruqui and Chris Dyer*. EACL 2014.
[Koehn 2005] **Europarl: A Parallel Corpus for Statistical Machine Translation**. *Philipp Koehn*. MT Summit 2005.
[McDonald et al. 2005] **Non-projective Dependency Parsing using Spanning Tree Algorithms**. *Ryan McDonald, Fernando Pereira, Kiril Ribarov and Jan Hajič*. EMNLP 2005.
[Nivre et al. 2007] **The CoNLL 2007 shared task on dependency parsing.** *Joakim Nivre, Johan Hall, Sandra Kubler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, Deniz Yuret.* EMNLP-CoNLL 2007
[Rasooli, Collins 2015] **Density-Driven Cross-Lingual Transfer of Dependency Parsers**. *Mohammad Sadegh Rasooli and Michael Collins.* EMNLP 2015.
[Täckström et al. 2013] **Target Language Adaptation of Discriminative Transfer Parsers**. *Oscar Täckström, Ryan McDonald, and Joakim Nivre.* NAACL 2013.
[Zhang, Barzilay 2015] **Hierarchical Low-Rank Tensors for Multilingual Transfer Parsing**. *Yuan Zhang and Regina Barzilay*. EMNLP 2015.