Multilingual Dependency Parsing for Low-Resource Languages via Multilingual Word Embeddings

KyungTae Lim (jujbob@gmail.com), Niko Partanen (niko.partanen@kotus.fi), Thierry Poibeau (Thierry.Poibeau@ens.fr)
https://github.com/jujbob/multilingual-bist-parser
https://github.com/jujbob/multilingual-models

Introduction

Dependency parsing for low-resource languages.

- This study presents a method for **parsing low-resource languages** with very small training corpora **using multilingual word embeddings** and annotated corpora of larger languages.
- The study demonstrates that specific language combinations enable improved dependency parsing when compared to previous work, allowing for wider reuse of pre-existing resources when parsing low-resource languages.

Contributions

A multilingual parsing approach with two contributions

- **New parser(s) for low-resource languages:**

We show that parsing performance can be improved by using additional resources (corpora and embeddings) for other languages.

- Specific language combinations enable improved dependency parsing
- Various tests conducted to evaluate different parsing scenarios (ongoing)

- **Building new resources:**

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau and Michael Rießler: The First Komi-Zyrian Universal Dependencies Treebanks. Universal Dependencies Workshop 2018.

- Two Komi-Zyrian UD corpora
- Several multilingual word embeddings with Komi-Zyrian and North Saami

Approach

Graph-based parsing with multilingual feature representations

- The parser learns from many languages, with several training data sets **including lexicalized features**.
- The parser adapts multilingual embedding for low-resource languages to improve parsing performance.

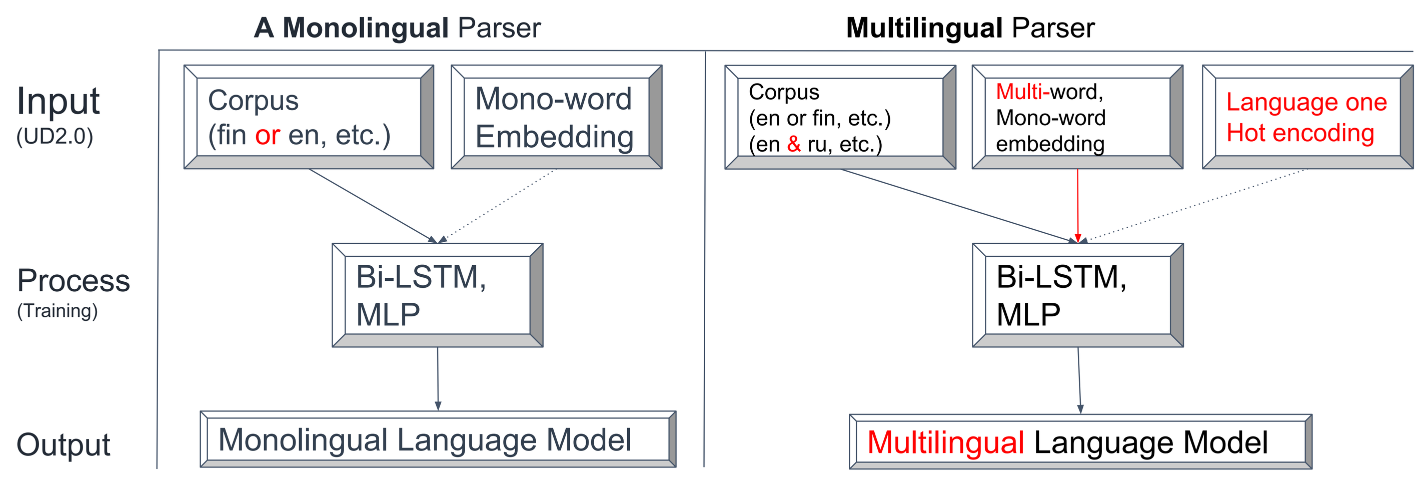


Figure 1: Example of differences between monolingual and multilingual approaches

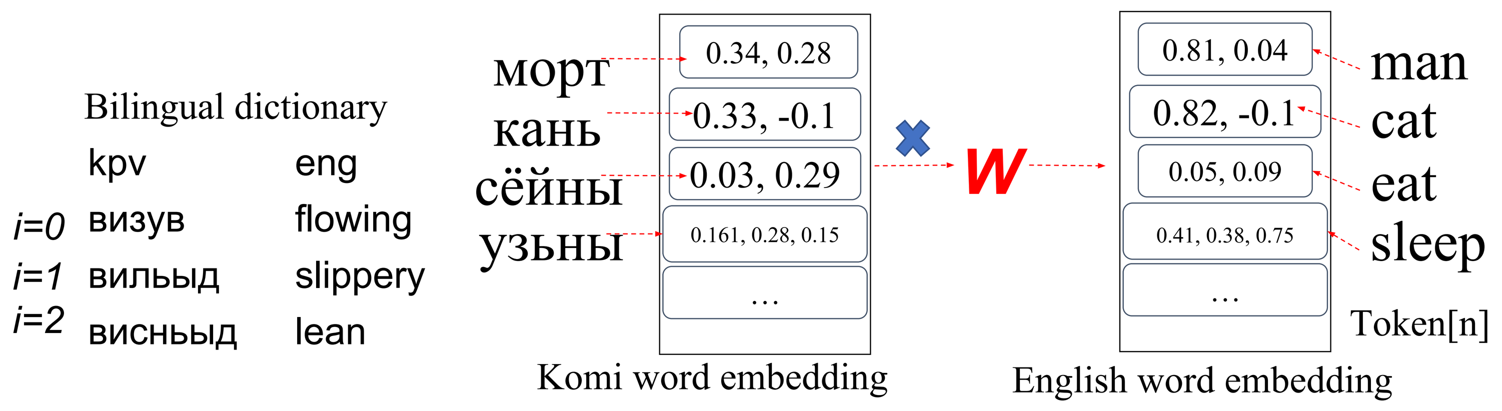
Multilingual Feature Transformation

Building multilingual word embedding

- **Finding a linear matrix that is minimizing the distance between embeddings:**

- Let X and Y be the source and target word embedding matrix so that xi refers to the ith word embedding of X and y_j refers to the jth word embedding of Y. And let D be a binary matrix, where Dij = 1, if x_i and y_j are aligned. Our goal is then to find a transformation matrix W such that Wx approximates y. This is done by minimizing the sum of squared errors:

$$\arg\min_{W} \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} \|x_i W - y_i\|^2$$



- Let us have a bilingual word embedding for English and Komi that is projected by ***W*** in a single vector space model. The distance of terms that have a similar meaning in English and Komi would be close.
- E.g. the cosine similarity between *"dog"* and *"пон"* is relatively high.

Overall structure

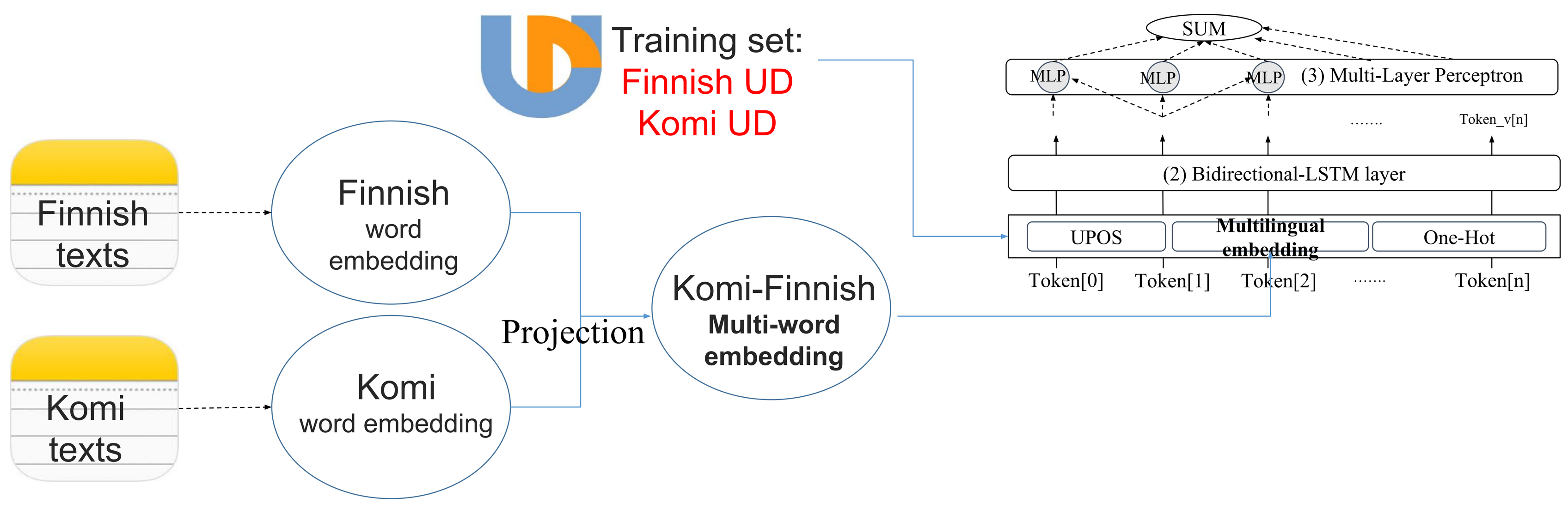


Figure 2: Example of Monolingual and Multilingual embeddings for Komi and Finnish

Training and Results

Evaluation results of Saami with 20 training sentences.

Case	Training corpus	LAS	UAS
1	sme (20)	32.96	46.85
2	eng (12,217)	32.72	50.44
3	fin (12,543)	40.74	54.24
4	sme (20) + eng (12,217)	46.54	61.61
5	**sme (20) + fin (12,543)**	**51.54**	**63.06**

Table 1: Labeled attachment scores (LAS) and unlabeled attachment scores (UAS) for North Saami (sme)

Corpus	Projected languages	UAS	LAS
hy_armntdp	Greek	1	1
br_keb	English	3	5
bxr_bdt	Russian	3	4
fo_oft	English	9	17
kk_ktb	Turkish	15	9
kmr_mg	English	3	4
pcm_nsc	-	21	18
sme_giella	Finnish+Russian	1	1
th_giella	English	21	21
hsb_ufal	Polish	2	2

Table 2: Languages trained with multilingual word embeddings and their ranking.

Evaluation results of Komi with 10 training sentences.

Bilingual pairs	Bi-dictionary	Bi-embedding
Finnish–Komi	12,879	2.3GB
Finnish–North Saami	12,398	2.4GB
Komi–English	8,746	7.5GB
North Saami–Finnish	10,541	2.4GB
Russian–Komi	12,354	5.7GB

Table 3: Dictionary sizes and size of bilingual word embeddings generated by each dictionary.

Case	Training corpus	LAS	UAS
1	kpv (10)	22.33	51.78
2	eng (12,217)	44.47	59.29
3	rus (3,850)	53.85	71.29
4	fin (12,543)	48.22	66.98
5	kpv (10) + eng (12,217)	50.47	66.23
6	kpv (10) + rus (3,850)	53.1	69.98
7	kpv (10) + fin (3,850)	53.66	71.29
8	kpv (10) + fin (12,543)	55.16	**73.73**
9	kpv (10) + eng (12,217) + fin (12,543)	52.5	68.57
10	**kpv (10) + rus (3,850) + fin (12,543)**	**56.66**	71.86

Table 4: Labeled attachment scores (LAS) and unlabeled attachment scores (UAS) for Komi-Zyrian (kpv).

KyungTae Lim, Niko Partanen and Thierry Poibeau: Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian. LREC 18.

Evaluation results for **Code-Switching** data

file	corpus	kpv	mixed	rus
kpv-ud-test.conllu	written monolingual	96.2%	3.8%	-
kpv-ud-test-mixed.conllu	written artificially mixed	70.2%	2.3%	27.5%
kpv-ud-ikdp.conllu	spoken	50.2%	9.9%	39.9%

Corpus	LAS	UAS
Written corpus	51.34	67.73
Artificially mixed corpus	53.61	65.74
Spoken corpus	54.77	68.20

Niko Partanen, Kyungtae Lim, Michael Rießler, Thierry Poibeau: Dependency Parsing of Code-Switching Data with Cross-Lingual Feature Representations. Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. p. 1–17.

Upcoming work

Better resources: training data, word lists, new parser

- **Earlier experiments should be repeated and refined:**
- More language pairs and combinations, more attention to minimizing differences in the resources used
- The SEx BiST (Semantically EXtended Bi-LSTM) parser performed well in the CoNLL 2018 Shared Task with North Saami
- New comparable word lists have became available in 2018
- Larger Komi-Zyrian treebanks enable broader training and testing

- Further questions:
  - Related languages and contact languages as resources in low-resource scenario
  - Taking better into account typological and contact-induced similarities
  - Better ways to evaluate the results and the exact relevance of different resources

Corpus	Projected languages	UAS	LAS
hy_armntdp	Greek	1	1
br_keb	English	3	5
bxr_bdt	Russian	3	4
fo_oft	English	9	17
kk_ktb	Turkish	15	9
kmr_mg	English	3	4
pcm_nsc	-	21	18
sme_giella	Finnish+Russian	1	1
th_giella	English	21	21
hsb_ufal	Polish	2	2

Table above originally published as Table 2 on page 148 in:

KyungTae Lim, Cheoneum Park, Changki Lee and Thierry Poibeau: SEx BiST: A Multi-Source Trainable Parser with Deep Contextualized Lexical Representations. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. p. 143–152.