A Semi-universal Pipelined Approach to the CoNLL 2017 UD Shared Task

Hiroshi Kanayama **Masayasu Muraoka** **Katsumasa Yoshikawa**
IBM Research - Tokyo
{hkana,mmuraoka,katsuy}@jp.ibm.com

Abstract

This paper presents the TRL team's system submitted for the CoNLL 2017 Shared Task, "Multilingual Parsing from Raw Text to Universal Dependencies." We ran the system for all languages with our own fully pipelined components without relying on either pre-trained baseline or machine learning techniques. We used only the universal part-of-speech tags and distance between words, and applied deterministic rules to assign labels. The delexicalized models are suitable for cross-lingual transfer or universal approaches. Experimental results show that our model performed well in some metrics and leads discussion on topics such as contribution of each component and on syntactic similarities among languages.

1 Introduction

We tested dependency-based syntactic parsing in 49 languages on Universal Dependencies (Nivre et al., 2015) using 81 corpora from the UD version 2.0 datasets (Nivre et al., 2017). The task is described in the overview paper (Zeman et al., 2017) and the whole system is evaluated on the TIRA platform (Potthast et al., 2014).

Instead of merely pursuing higher scores in the shared task, we adopted several strategies in the design of our parser:

Self-contained system. To keep capabilities to control the input and output of the system, we use only our own components for the whole pipeline including sentence splitter, tokenizer, lemmatizer, PoS tagger, dependency parser and role labeler. We do not rely on any existing preprocessors such as UDPipe (Straka et al., 2016) and SyntaxNet (Weiss et al., 2015).

One model per language. When there are multiple corpora in a language with different annotation strategies, our system does not optimize models for each corpus, because the real applications do not assume such specific corpora.

No machine learning. We use merely simple statistics with parts of speech of each word and distance between words, and induced deterministic rules. Neither higher order models nor word embeddings are used, thus our system is fully controllable with linguistic knowledge.

Componentized pipeline. Components in the pipeline can be divided and optimized independently so that they are interchangeable with other corresponding components such as the UDPipe tokenizer. Our dependency parser relies only on Universal PoS tags and does not use an extended PoS, lemma nor features annotated by a specific tokenizer.

Our system was composed under these constraints at the sacrifice of overall scores but it performed marginally well, achieving the best participant scores in a number of metrics. The major contributions in this report are as follows:

1. Report of runs without UDPipe with very different results than those obtained from other participants.
2. Experiments in cross-lingual and universal scenarios by using delexicalized statistics of different languages.
3. Simple and reusable techniques to induce rules for PoS tagging and relation labeling.

Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 265–273,
Vancouver, Canada, August 3-4, 2017.

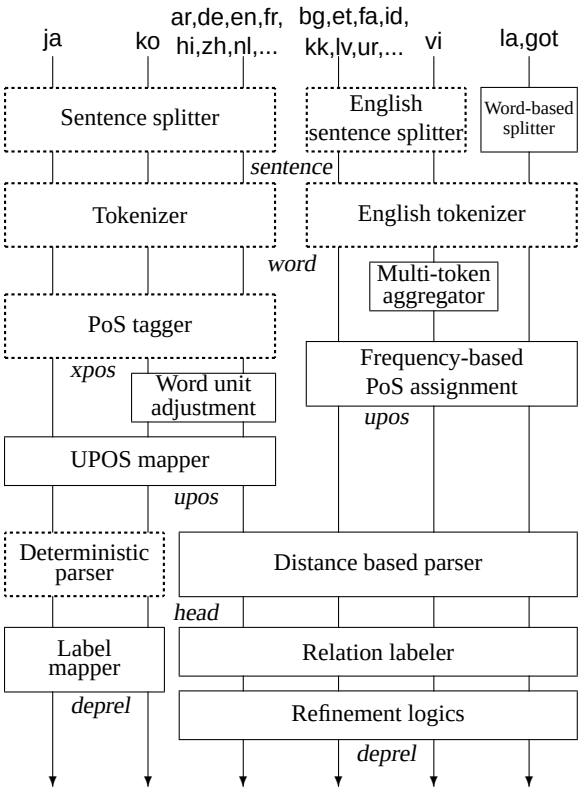

Figure 1: The pipelined architecture for multilingual parsing from raw text. Dotted boxes indicate existing (not UD-compliant) components.

Section 2 describes each components in our pipeline. Section 3 reports our results, including ablation studies and additional experiments in cross-lingual and multilingual settings. Section 4 shows some related prior work related to our approach.

2 Components

Figure 1 illustrates our pipelined architecture for multilingual parsing from raw text. As indicated as dotted boxes in the figure, we exploited in-house engines for sentence splitting, tokenization and PoS tagging for a number of languages and fit them to the UD annotation schemata. For languages which our engine does not cover, we used simple statistics in the training corpus to assign Universal PoS (UPOS). For syntactic parsing, we extracted statistics to predict the head words, taking into account UPOS and distance. To assign relation labels we applied rules induced from the corpus.

The rest of this section describes each component with language specific treatments in the order in the pipeline.

2.1 Sentence splitting

For the sentence splitting we applied existing logics, taking into account language specific punctuations and special cases such as “Mr.” in English. For languages that our sentence splitter does not cover, we simply applied the logic for English. For corpora that do not use punctuation at all (*e.g.* got and la_proiel), we identified words that tend to be the first or the last word in a sentence (more than half of appearance *e.g.* “itaque” in Latin as the first word), and used them to split long sentences that had 10 or more words.

2.2 Tokenization

Our in-house engine tokenizer and PoS tagger support 17 languages; ar, cs, da, de, en, es, fr, he, it, ja, ko, nl, pl, pt, tr, ru and zh. For three of them, Japanese (ja), Korean (ko) and Chinese (zh), words are split in very different manner without relying on white spaces [1].

We applied English tokenizer for other languages to simply split words by white spaces and punctuations. For Vietnamese (vi) in which the word units are longer than space-split tokens, we extracted multi-token words from the training corpus and aggregated them in runtime. This raised the word F1 score for vi from 73.7 to 85.1.

There are unignorable mismatches in tokenization strategies between our tokenizer and UD corpora. The major difference is in Korean (ko): while our tokenizer splits particles and suffixes from content words, the UD corpus gives whitespace (*eojeol*) tokenization. Accordingly, we merged those tokens after getting parts of speech of each unit.

We also made adjustments in Turkish (tr) to attach suffixes except for “ki”, and in Arabic (ar) to attach the determiner “al”. There still remains many differences in other languages but we did not make any other modifications, which resulted in lower word correspondence values (95.5 on average) compared to those of UDPipe (98.6).

2.3 PoS tagging

As well as the tokenization, we applied PoS tags output by our engine for 17 languages to get their own PoS schema; some of them are close to the Penn Treebank style and the others are in different schemes. We adopted those tags as Extended

[1] Though the word unit in the Korean corpus in UD2.0 is determined by white space, our tokenizer gives finer tokens by splitting functional words.

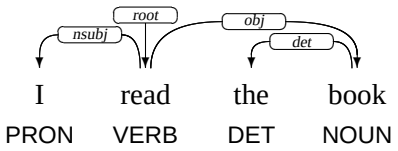

Figure 2: A sample dependency structure of an English sentence.

PoS (XPOS) tags and mapped them to UPOS. The mapper assigns the most frequent UPOS in the training corpus for a combination of XPOS and the lemma of a given word.

By definition, our PoS tagger does not distinguish some of the main verbs (VERB) from auxiliary verbs (AUX) such as "do" and "have" in English, "avoir" in French and "haber" in Spanish, which causes many parsing errors, and so we added heuristics to change the UPOS using the context.

For other languages the PoS tagger does not cover, we assigned the most frequent UPOS for each surface form in the training corpus. Even with this naïve method we obtained UPOS scores higher than 90 for some languages such as Czech (cs), Persian (fa), Hindi (hi) and Indonesian (id) but it did not work well enough for lower resource languages.

2.4 Dependency parsing

2.4.1 PoS-level models

To keep the simplicity and language universality of the parsing method, we built the first-order delexicalized model for each language[2]. The score of the dependency between two words is determined only by the UPOS of head and dependent words and surface distance between two words.

Figure 2 shows a sample dependency structure for an English sentence and Table 1 shows true (T) and false (F) dependencies found in the sentence in Figure 2. By counting frequencies of these events for all pairs in a sentence, the ratio of correct dependency for a pair of PoS and distance is calculated.

Formally, let h be a head word, d be a dependent, p_w be the UPOS of w, and $\Delta_{d,h}$ be the distance[3] between d and h, so that the score is

[2]Not for each corpus, following 'one model par language' policy.

[3]The difference of word IDs of h and d. We cap the maximum distance at 12 (empirically determined), *i.e.* word pairs further than 13 are regarded as $\Delta = 12$.

dependent	head	distance	dependency?
PRON	VERB	1	T
PRON	DET	2	F
PRON	NOUN	3	F
VERB	PRON	-1	F
VERB	DET	1	F
VERB	NOUN	2	F
DET	PRON	-2	F
DET	VERB	-1	F
DET	NOUN	1	T
NOUN	PRON	-3	F
NOUN	VERB	-2	T
NOUN	DET	-1	F

Table 1: True (T) and false (F) dependencies between two words in the sentence in Figure 2. Negative distance means that the head is left to the dependent.

English (en)	
ADJ, NOUN, -1	.238
ADJ, NOUN, 1	.906
ADJ, NOUN, 2	.639
VERB, ADJ, -2	.512
ADP, NOUN, 2	.817
AUX, ADP, 1	.034

French (fr)	
ADJ, NOUN, -1	.959
ADJ, NOUN, 1	.967
ADJ, NOUN, 2	.130
VERB, ADJ, -2	.180
ADP, NOUN, 2	.943
AUX, ADP, 1	.000

Table 2: Examples of dependency scores between two words for English and French. A condition indicates the PoS of dependent and head words, and distance between two words.

$$\frac{\#(\mathrm{T} \mid p_d, p_h, \Delta_{d,h})}{\#(\mathrm{T} \mid p_d, p_h, \Delta_{d,h}) + \#(\mathrm{F} \mid p_d, p_h, \Delta_{d,h})},$$

where $\#(\cdot)$ is the frequency in the training corpus, T denotes that d depends on h and F denotes it does not. The score is set to 0 when the denominator is 0.

Table 2 shows example scores. These statistics reflect universal attributes, for example, smaller distance is preferred, functional words tend not to have dependents, and so on. Also language specific attributes are contained, such as regarding orientation of adjective modification and adpositions.

These scores are used as the weight of the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to obtain the minimum spanning tree to optimize the dependency structures in a

sentence. This algorithm can produce a non-projective tree, which frequently appears in languages such as German, Latin and Czech (McDonald et al., 2005).

2.4.2 Language specific cases

Japanese (ja) and Korean (ko) are parsed in a different manner. A common point to both languages is that all content words form right-head structures; consequently, a set of rules selects the syntactically possible head words for a given word by using the syntactic features (Kanayama et al., 2014). Here the dependencies are determined as the nearest baseline among the modification candidates without relying on the statistics of the training corpora.

For 'surprise languages' that do not have training corpora, we use models for languages in the close regions (Russian (ru) for Buryat (bxr), Persian (fa) for Kurmanji (kmr), Finnish (fi) for North Sámi (sme) and Polish (pl) for Upper Sorbian (hsb)) but these selections were not optimal as found in the experiments in Section 3.2.

2.4.3 Exceptional dependencies

The statistic model above is apparently ignorant of the vocabulary and lexical features finer than UPOS level. To capture some phenomena we made two deterministic modifications.

Fixed expressions. Multi-word expressions behave exceptionally in the UPOS-based model. In each language we extracted fixed phrases such as "because of" and "as well as" in English, and in runtime forcibly tagged dependencies for such word sequences with '*fixed*' label. Also, for consecutive appearance of same PoS tags of NOUN, PRON or NUM, a structure with the majority label (one of *flat*, *nmod*, *compound* or *nummod*) is assigned depending on the pairs of language and PoS, *e.g.* give left-head structures with *flat* label for PROPN sequences of Catalan (*ca*).

Consistent words. English UPOS PART is used for possessive "'s" and infinitive "to", which behaves very differently from each other. For such words whose head word is in a consistent direction per dependent word, the score for the other direction is discounted by multiplying 0.1, *e.g.* 0.1 is multiplied for the score of left-head modification of PART: "to" in English.

English (en)	
ADP, NOUN, +	*case*
VERB, NOUN,+	*acl*
NOUN, VERB, +	*nsubj*
NOUN, VERB, −	*obj*
ADJ, VERB, −	*xcomp*

Russian (ru)	
ADP, NOUN, +	*case*
VERB, NOUN,+	*amod*
NOUN, VERB, +	*nsubj*
NOUN, VERB, −	*obl*
ADJ, VERB, −	*obl*

Table 3: Examples of label assignment for English and Russian. '+' and '−' indicate the direction of the head word against a dependent; '+' means that the head comes right to a dependent.

2.5 Relation label assignment

After getting the tree structures, we assigned dependency labels to each node by referring to the most frequent label between two UPOS tags in the languages. The labels vary by language and orientation of the dependencies as exemplified in Table 3.

In some cases the labels are difficult to deterministically assign merely by using UPOS of two words. In such cases, we applied the following label refinement rules.

Word based constraints. Forcibly change the label for words whose relation labels are mostly consistent ($\geq .95$), *e.g.* modification by "there" in English should have *expl* label.

Verb arguments. Adjust the label of NOUN, PROPN and PRON when the word is a dependent of VERB with several conditions, *e.g.* set *obl* if the word has a dependent labeled *case* in most of languages.

Pronouns. Change the relation label of PRON as a dependent of VERB to its majority[4] for a surface form. *E.g.* select *obj* for "him" in English.

Conjunctions. When the dependent and head words have the same UPOS and there is CCONJ between the two words, set the label of the dependent as *conj*.

[4] Among *nsubj, obj, iobj* and *expl.*

	Submitted results (without UDPipe)						UDPipe preprocess					
Language	Sentence	Words	UPOS	UAS	LAS	WLAS	Sentence	Words	UPOS	UAS	LAS	WLAS
∗ Average	79.99	95.47	80.45	53.53	43.37	37.33	88.48	98.61	91.02	61.89	52.12	45.75
ar	77.10	92.21	80.05	51.44	39.98	33.40	84.57	93.69	88.13	55.26	46.41	40.53
ar_pud	99.10	**96.05**	73.06	56.59	42.76	35.61	100.00	90.82	70.27	50.05	39.51	33.98
bg	80.71	97.43	88.27	61.26	53.39	46.33	92.83	99.91	97.58	74.43	68.00	63.39
bxr	**93.69**	98.44	47.69	23.40	14.02	4.22	91.81	99.35	84.12	40.35	25.97	19.22
ca	97.25	91.96	85.90	58.31	51.43	40.77	98.95	99.97	98.04	72.85	66.44	56.58
cs	77.90	97.13	93.29	61.32	54.81	51.92	92.03	99.90	98.13	67.24	60.15	57.31
cs_cac	99.76	97.45	92.90	64.88	57.59	53.82	100.00	99.99	98.27	69.40	61.93	57.99
cs_cltt	45.04	88.90	79.91	54.09	48.38	46.30	95.06	99.35	95.41	63.20	56.50	53.48
cs_pud	87.57	97.68	92.67	64.47	58.02	55.20	96.43	99.29	96.55	66.44	59.78	57.34
cu	1.16	99.97	85.52	51.09	35.00	29.90	36.05	99.96	93.34	60.42	45.77	41.04
da	72.76	93.61	79.57	48.62	41.01	34.67	79.36	99.69	95.04	63.15	55.91	50.64
de	69.84	89.86	79.32	49.28	43.23	37.93	79.11	99.65	90.83	62.17	55.87	47.83
de_pud	86.93	93.01	77.55	51.35	43.44	38.46	86.49	98.00	84.38	61.76	53.17	46.38
el	81.03	98.12	86.66	61.06	52.44	41.08	90.79	99.88	95.18	72.34	66.77	59.09
en	64.77	94.31	82.41	56.44	49.56	44.77	73.22	98.67	93.11	62.83	55.98	50.41
en_lines	81.72	98.63	86.90	59.77	52.49	47.25	85.84	99.94	94.53	66.12	57.65	52.00
en_partut	90.61	99.45	88.32	62.48	54.50	46.56	97.51	99.49	93.03	65.19	57.31	49.48
en_pud	94.29	99.04	85.43	60.05	53.94	48.09	97.13	99.66	94.00	66.27	59.82	53.30
es	88.19	96.13	88.82	63.80	57.01	49.03	94.15	99.69	95.60	69.76	63.55	53.34
es_ancora	95.44	98.54	89.35	65.22	55.52	45.81	97.05	99.95	98.15	70.42	60.92	51.31
es_pud	95.40	96.85	84.45	66.68	58.40	50.03	93.42	99.47	88.15	71.92	64.19	54.46
et	80.91	98.73	79.68	49.79	34.05	29.83	85.20	99.77	87.70	58.20	43.25	39.60
eu	91.06	97.52	85.58	53.63	41.80	36.35	99.58	99.96	92.36	59.97	47.51	41.92
fa	96.11	99.07	92.32	56.28	48.40	41.61	98.00	99.64	96.00	60.23	52.12	45.25
fi	80.34	96.10	84.02	45.45	31.51	32.98	84.56	99.63	94.01	51.39	38.23	40.40
fi_ftb	75.42	98.66	78.95	57.45	44.55	34.06	83.83	99.88	91.87	63.80	52.65	42.04
fi_pud	87.50	96.28	85.59	46.67	32.55	35.19	93.67	99.61	95.61	52.92	39.68	43.12
fr	86.99	93.78	85.75	61.71	54.59	50.00	93.59	98.87	95.33	70.25	64.06	57.61
fr_partut	89.99	94.43	86.64	64.75	57.18	51.92	98.00	98.95	94.46	69.93	62.90	54.73
fr_pud	95.85	95.01	81.19	64.04	56.96	51.35	92.32	98.17	87.90	69.42	63.22	56.96
fr_sequoia	67.24	92.70	84.35	59.60	52.75	48.63	83.75	99.06	95.40	70.50	64.35	58.49
ga	95.49	96.44	82.54	61.87	43.80	28.38	95.81	99.29	88.17	65.09	48.30	32.66
gl	90.64	98.39	88.73	64.69	55.70	45.49	96.15	99.92	96.84	68.78	62.96	54.76
gl_treegal	78.79	87.39	71.85	46.75	33.25	27.85	81.63	98.62	90.69	62.92	50.72	41.85
got	3.20	99.90	86.85	51.80	36.92	30.18	27.85	100.00	93.55	58.28	44.97	39.42
grc	41.91	99.96	71.46	41.18	28.63	18.68	98.43	99.95	82.13	47.68	36.16	28.55
grc_proiel	1.42	98.93	87.75	49.97	38.88	26.43	43.11	100.00	95.72	57.61	46.97	36.12
he	98.89	84.45	73.29	47.31	37.10	26.50	99.39	84.82	80.48	51.21	42.52	31.92
hi	90.22	99.06	90.15	66.60	55.15	40.59	99.20	100.00	95.63	71.92	60.45	45.90
hi_pud	94.47	**99.65**	81.48	51.26	36.80	26.09	90.83	97.81	83.75	52.57	39.16	29.68
hr	84.67	98.01	87.71	56.49	45.82	41.33	96.92	99.93	95.67	66.95	57.77	54.60
hsb	68.23	99.51	61.48	30.74	22.15	15.99	90.69	99.84	90.30	51.85	41.96	37.46
hu	88.86	94.42	78.98	46.56	33.75	28.71	93.85	99.82	90.80	61.85	49.66	45.60
id	85.37	99.07	90.54	63.72	54.97	51.56	91.15	99.99	93.32	65.36	57.37	54.87
it	89.05	88.56	77.66	57.96	49.67	46.57	97.10	99.73	97.07	76.45	70.78	61.96
it_pud	**97.81**	89.19	73.72	58.76	50.34	46.67	96.58	99.17	93.07	75.50	70.09	61.20
ja	94.56	**98.59**	**98.45**	**91.14**	**91.13**	**84.45**	94.92	89.68	98.54	61.28	58.54	43.95
∗ ja_pud	**97.42**	98.89	98.52	88.79	88.71	80.09	94.89	91.06	88.69	64.84	62.31	46.50
kk	**89.35**	95.93	56.39	**45.72**	24.14	18.30	81.38	94.91	50.06	31.53	17.01	13.11
kmr	**98.64**	96.86	35.55	10.59	3.44	3.67	97.02	98.85	90.04	47.66	35.31	28.97
ko	69.87	98.12	76.56	55.54	45.83	42.12	93.05	99.73	93.79	53.11	26.48	20.90
la	62.03	**100.00**	73.24	37.40	23.48	20.35	98.09	99.99	83.39	42.86	29.75	26.54
la_ittb	74.82	99.19	91.33	47.21	37.59	32.05	93.24	99.99	97.21	54.15	44.38	38.52
la_proiel	1.25	99.77	85.27	43.87	27.97	22.30	25.80	100.00	94.82	49.74	34.78	28.95
lv	97.36	98.07	78.66	45.13	34.21	29.09	98.59	98.91	88.37	53.17	43.25	38.83
nl	72.92	92.67	78.53	47.58	39.44	30.28	77.14	99.88	91.00	60.45	50.76	41.36
nl_lassysmall	33.65	93.68	75.54	45.95	35.32	26.90	78.62	99.93	96.86	63.07	52.78	45.58
no_bokmaal	86.72	95.51	87.63	57.07	49.31	42.90	95.76	99.75	96.75	69.57	62.13	55.81
no_nynorsk	80.12	94.88	86.43	54.20	46.83	40.78	91.23	99.85	96.38	66.32	58.91	52.00
pl	97.83	97.19	89.14	68.47	57.83	52.71	98.91	99.88	95.31	76.81	65.44	59.39
pt	77.79	86.46	71.60	52.04	40.46	35.71	89.79	99.52	96.22	72.50	62.11	50.03
pt_br	92.65	88.26	70.12	53.10	43.76	38.87	96.84	99.84	96.97	72.17	64.03	52.71
pt_pud	95.80	88.69	71.70	54.83	44.86	39.42	95.65	99.42	88.45	70.34	60.62	49.06
ro	89.13	96.37	87.64	64.24	53.35	46.51	93.42	99.64	96.40	70.76	62.20	55.63
ru	91.52	93.87	85.62	55.70	45.52	48.67	96.42	99.91	94.47	60.95	51.16	53.69
ru_pud	95.28	**98.29**	86.39	58.58	49.04	52.72	98.95	97.18	85.85	59.33	49.91	54.27
ru_syntagrus	89.75	98.36	89.44	69.49	61.66	51.91	97.81	99.57	97.99	72.58	65.84	56.42
sk	68.30	99.75	81.11	46.60	38.69	31.87	83.53	100.00	92.19	68.28	60.78	57.56
sl	96.49	99.73	86.89	56.55	47.25	39.89	99.24	99.96	96.34	72.11	64.63	60.02
sl_sst	0.52	88.87	77.78	38.65	29.92	24.63	16.72	99.82	88.82	50.20	39.82	35.27
sme	**99.13**	98.28	43.88	27.93	7.47	7.74	98.79	99.88	86.81	46.05	31.46	33.56
sv	89.69	94.51	82.37	54.98	45.17	39.76	96.37	99.84	95.41	67.40	59.48	54.79
sv_lines	79.82	96.59	83.09	57.70	47.13	42.15	86.44	99.98	94.22	68.18	60.47	55.21
sv_pud	**95.52**	95.00	80.36	54.05	42.58	36.82	90.20	98.26	91.16	65.35	56.49	51.29
tr	93.57	88.78	77.25	42.07	30.48	25.91	96.63	97.89	91.22	52.19	39.28	33.70
tr_pud	88.88	88.21	62.02	39.14	21.36	15.17	93.91	96.62	71.05	48.82	25.72	18.36
ug	69.05	98.23	65.27	48.10	23.72	14.37	63.55	98.52	73.63	50.40	31.20	22.55
uk	91.80	98.58	73.97	45.91	33.99	23.45	92.59	99.81	86.72	64.94	52.14	44.82
ur	97.93	97.70	86.31	62.47	50.26	36.55	98.32	100.00	91.71	69.35	56.90	42.73
vi	86.12	85.41	74.53	37.13	31.01	28.50	92.59	82.47	73.82	35.12	29.54	26.32
zh	92.81	83.64	71.31	31.49	25.60	23.24	98.19	88.91	82.69	39.65	33.87	30.99

Table 4: Overall F1 scores over test data (see Section 3.1).

3 Experimental Results

3.1 Overall results

Table 4 shows the results for 81 test corpora in 49 languages including 'surprise languages'.

The left side shows the performance of our system described in Section 2. The scores are the same as those in the official run except for ja_pud data on which we encountered a technical problem in the official run. '$*$' denotes that the values were updated from the official score. WLAS denotes "Weighted labeled attachment score", which discounts the functional word attachments by multiplying 0.1 and ignores punctuation.

Numbers in bold letter indicate that our system achieved the best scores among task participants. Our sentence splitting was the best for seven corpora including three surprize languages, and word segmentation was best for five corpora.

Japanese (ja) shows the best score except for sentence splitting[5], but it is exceptional here. As we provided the Japanese UD2.0 data set, we have the consistent tokenization, PoS mapping and label definition with the data set, thus it is straightforward to convert the parsing structure into appropriate UD schema. We intentionally use the naïve method for parsing (nearest baseline), however, we performed the best among the participants due to the high coincidence of the tokenization.

For Kazakh (kk), our approach worked well and achieved the best score in sentence splitting and unlabeled attachment scores (UAS). The absolute score was not high, so this shows the difficulty of the language for machine learning approaches.

Besides the difficult languages in terms of tokenization: Chinese (zh), Vietnamese (vi) and Hebrew (he), some languages show quite low scores for word splitting (*e.g.* pt and tr) due to differences in tokenization policies which our adjustment rules did not cover. Due to the nature of the pipelined architecture, the errors in word splitting directly affect the downstream metrics. Since the UPOS is used for dependency parsing, PoS tagging and PoS mapping errors are critical for parsing scores, both UAS and LAS.

The right side of Table 4 shows the results of our parser using UDPipe for tokenization and PoS tagging. Three columns (Sentence, Words and UPOS) show the scores of UDPipe itself, and the rest of columns show the scores of our parser when UDPipe was applied for preprocessing. Since UDPipe was trained with the UD corpora the tokenization and PoS tagging performed much better than ours and resulted in scores 8 and 9 points higher than those obtained for UAS and LAS respectively. For Kazakh (kk), Vietnamese (vi) and one of Arabic data (ar_pud), our tokenizer and PoS tagger performed better than UDPipe, resulting in better parsing scores in the left side.

Korean (ko) is another exception. Since our UPOS-based dependency parsing model does not capture the decomposed elements of each token, the parser did not work well after the UDPipe preprocessing. Our deterministic parser can handle the functional words thus it performed better.

3.2 Cross-lingual and universal evaluation

One of the advantages of Universal Dependencies is the capability to test the language independent model and cross-lingual transfer learning. As described in Section 2.4, our dependency parsing models without any lexical information are very general. They therefore can be applied to other languages enabling us to test a universal language model.

Table 5 compares the UAS scores with the cross-lingual and universal settings. The 'Own model' column shows the original score, the 'Best transfer' column shows the score using the model that performed the best among different languages, and the 'Universal' column shows the score obtained with the combined statistics extracted from all of the multilingual corpora. Numbers in **bold** denote that the transfer or universal model outperformed the language specific model. Japanese (ja) and Korean (ko) were not tested here because they did not use compatible models.

The experimental result shows the best models for applying low-resource languages: fi for bxr, cs for kmr, tr for sme and hr for hsb. Also for relatively low-resource languages such as Kazakh (kk) and Ukrainian (uk), the models with larger corpora outperformed their own models. For four French (fr) corpora, the Portuguese (pt) model performed as well as the French model. This suggests the model with three different French corpora generated a noisy model.

It is interesting to consider the 'neighbor' languages in terms of syntax. English and Swedish (sv) selected each other as the closest languages, which suggests that they are selected not only be-

[5]Interestingly the sentence splitting score is almost the lowest among participants.

Language	Own model	Best transfer		Universal
Average	53.53			49.06
ar	51.44	ga	47.45	45.47
ar_pud	56.59	ga	55.92	54.53
bg	61.26	cs	60.38	60.19
bxr	23.40	fi	**24.66**	18.10
ca	58.31	es	57.90	57.57
cs	61.32	sl	60.68	59.31
cs_cac	64.88	sl	64.19	62.93
cs_cltt	54.09	sl	52.88	52.34
cs_pud	64.47	sl	64.06	62.43
cu	51.09	got	51.06	47.29
da	48.62	no	48.45	47.56
de	49.28	nl	47.19	47.18
de_pud	51.35	sl	49.92	49.21
el	61.06	de	57.04	58.37
en	56.44	sv	54.84	52.15
en_lines	59.77	sv	58.89	56.05
en_partut	62.48	sv	61.18	58.60
en_pud	60.05	sv	58.55	56.00
es	63.80	pt	63.54	62.51
es_ancora	65.22	ca	64.97	63.96
es_pud	66.68	pt	66.34	65.37
et	49.79	en	48.57	44.14
eu	53.63	hu	47.05	42.78
fa	56.28	la	52.28	51.92
fi	45.45	en	43.98	39.58
fi_ftb	57.45	en	50.92	50.51
fi_pud	46.67	en	45.23	41.83
fr	61.71	pt	61.48	60.18
fr_partut	64.75	it	**64.98**	63.55
fr_pud	64.04	pt	**64.65**	63.56
fr_sequoia	59.60	pt	59.24	57.74
ga	61.87	ro	59.05	57.05
gl	64.69	pt	62.62	60.99
gl_treegal	46.75	pt	**48.64**	**47.67**
got	51.80	grc	50.93	49.22
grc	41.18	no	39.13	39.75
grc_proiel	49.97	got	48.05	46.70
he	47.31	pt	45.35	43.83
hi	66.60	ur	65.48	38.36
hi_pud	51.26	ur	50.83	33.41
hr	56.49	cs	55.36	54.23
hsb	30.74	hr	**32.37**	**32.06**
hu	46.56	fi	46.21	36.04
id	63.72	ro	62.58	61.55
it	57.96	pt	**58.19**	57.09
it_pud	58.76	pt	**59.38**	58.75
ja	91.14	-		-
ja_pud	88.79	-		-
kk	45.72	ur	**47.58**	18.77
kmr	10.59	cs	**12.32**	**12.49**
ko	55.54	-		-
la	37.40	grc	**39.78**	35.83
la_ittb	47.21	cs	45.01	45.19
la_proiel	43.87	got	42.17	41.94
lv	45.13	en	44.80	37.46
nl	47.58	de	45.87	45.65
nl_lassysmall	45.95	la	44.10	43.68
no_bokmaal	57.07	da	55.69	54.70
no_nynorsk	54.20	sv	52.60	51.64
pl	68.47	cs	66.65	66.75
pt	52.04	ca	51.71	50.32
pt_br	53.10	es	53.08	51.56
pt_pud	54.83	es	**54.96**	53.12
ro	64.24	pt	63.60	62.62
ru	55.70	bg	**58.63**	**57.12**
ru_pud	58.58	cs	**64.39**	**62.29**
ru_syntagrus	69.49	fi	58.81	59.63
sk	46.60	cs	**48.82**	**47.45**
sl	56.55	cs	56.21	54.39
sl_sst	38.65	hr	37.35	37.78
sme	27.93	tr	**28.73**	20.31
sv	54.98	en	54.20	51.70
sv_lines	57.70	en	56.93	54.04
sv_pud	54.05	en	53.99	50.47
tr	42.07	ug	37.71	28.81
tr_pud	39.14	ug	36.85	24.12
ug	48.10	ur	45.41	16.20
uk	45.91	sv	**48.14**	**46.15**
ur	62.47	hi	60.35	32.83
vi	37.13	en	33.36	34.15
zh	31.49	tr	27.17	19.58

Table 5: UAS-F1 scores with language specific models, and transfer models (see Section 3.2.

cause of the size of the training corpora. It is also notable that two variants of Norwegian (no) were closest for different languages (Danish (da) and Swedish).

Even the universal model performed well. The drop in UAS scores from the language-specific result was only 4.5 points on average. This shows our method is general enough for multilingual design. Not only for low-resource languages such as Ukrainian and surprise languages, but also for Russian (ru) and Slovak (sk), the universal model outperformed the language specific model.

3.3 Ablation of refinement rules

Table 6 shows the difference in UAS scores when we did not apply one of the sets of rules to change the dependency structures described in Section 2.4 and LAS scores without one of refinements for relation labels described in Section 2.5. The identification of multi word tokens did not work well as expected, and the word level rules made little contribution.

Applying all label refinement rules improved the LAS score by 2.35 points on average. The rules to modify labels for verb arguments were the most important on average. Conjunction rules were very simple but consistently improved for almost all languages. Word-based constraints are good for some languages but may cause side effects. Pronoun rules were good for Gothic (got), which suggests that the Gothic pronouns are relatively consistently used for argument cases *e.g.* "saei" for *nsubj* and "mik" for *obj*.

4 Related Work

Some approaches share the same motivation with ours. Martínez-Alonso et al. (2017) used a small set of UPOS-level attachment rules for parsing and achieved 55 UAS with a universal model with predicted PoS. In this shared task we need to tackle the preprocessing and relation labeling as well which cannot be done in a language agnostic manner. Accordingly, we used minimum statistics for each language and achieved UAS levels similar to those for our own tokenization and PoS prediction, and higher value (by 9 points) when we use the UDPipe preprocessor.

Universal parsing is not our main focus here, but our results in the rightmost column in Table 2 can be used to compare our approach with universal approaches (Ammar et al., 2016).

	ΔUAS		ΔLAS				
Language	fixed	const	word	arg	pron	conj	ALL
Average	-0.35	0.06	0.41	1.17	0.18	0.40	2.35
ar	-2.41	0.04	1.51	-0.26	0.00	0.29	1.54
ar_pud	-1.39	0.00	1.44	2.80	-0.01	0.68	4.85
bg	-0.54	0.10	0.88	0.88	0.58	0.73	3.62
bxr	-3.32	0.00	0.00	0.00	0.00	0.00	0.00
ca	1.18	0.05	0.08	0.24	0.49	0.59	1.54
cs	-0.39	0.06	0.24	1.95	0.21	0.50	3.42
cs_cac	-0.13	0.07	0.14	1.93	0.11	1.17	3.72
cs_cltt	0.67	0.06	0.13	0.98	0.02	1.27	2.55
cs_pud	-0.14	0.09	0.08	2.43	0.19	0.60	3.97
cu	0.00	0.16	2.36	2.14	0.49	0.53	6.46
da	0.53	0.04	0.09	1.50	0.01	0.24	2.22
de	-0.01	0.01	0.43	1.56	0.27	0.50	2.94
de_pud	0.00	0.01	0.12	1.77	0.44	0.38	2.91
el	0.04	0.03	1.60	1.55	0.02	0.63	3.77
en	0.04	0.07	0.52	1.57	-0.33	0.40	2.16
en_lines	0.19	0.05	0.59	2.14	-0.35	0.38	2.76
en_partut	0.12	0.03	0.33	1.90	-0.04	0.82	3.00
en_pud	0.02	0.06	0.32	2.12	-0.11	0.48	2.82
es	-0.15	0.05	-0.10	2.14	0.53	0.59	4.27
es_ancora	0.41	0.00	-0.63	0.68	-0.08	0.43	-0.46
es_pud	0.21	0.02	0.20	2.79	0.32	0.54	3.93
et	0.08	0.00	1.14	0.29	0.23	0.05	1.83
eu	-0.16	0.16	1.29	0.37	0.01	0.02	1.62
fa	-4.35	0.19	0.82	1.44	0.00	1.49	3.93
fi	0.13	-0.07	-0.84	0.36	0.14	0.04	-0.29
fi_ftb	0.21	0.12	-0.33	-0.09	0.20	0.07	-0.09
fi_pud	0.09	-0.24	-0.65	0.48	0.10	0.10	0.05
fr	-1.08	0.11	-0.37	2.37	0.40	0.48	4.11
fr_partut	-2.64	0.01	-0.69	2.26	0.24	0.72	3.75
fr_pud	-1.10	0.06	-0.11	2.81	0.40	0.40	4.98
fr_sequoia	0.16	0.05	-0.58	1.75	0.16	0.33	2.95
ga	0.53	0.32	0.46	0.08	0.05	0.44	1.08
gl	-0.45	0.04	0.34	0.99	-0.04	-0.41	0.88
gl_treegal	0.02	0.08	0.28	0.29	0.11	0.80	1.50
got	-0.36	0.15	0.69	1.20	2.18	0.54	4.69
grc	-0.12	0.02	-1.02	1.13	0.12	0.47	0.71
grc_proiel	-0.37	0.01	3.13	1.50	1.07	0.47	6.24
he	0.10	0.09	0.41	0.60	0.02	0.36	1.38
hi	0.03	0.02	0.25	0.95	0.05	-0.01	1.61
hi_pud	0.00	0.03	0.74	0.40	0.27	0.04	1.84
hr	-0.59	0.00	0.75	1.41	0.24	0.81	3.46
hsb	-0.32	0.00	0.29	0.00	0.00	0.66	0.96
hu	-0.07	0.07	0.29	0.26	0.04	0.01	0.64
id	-2.53	0.05	0.05	2.43	-0.03	0.76	3.21
it	0.02	0.06	-0.14	1.99	0.11	0.52	3.13
it_pud	0.16	0.08	-0.01	2.60	0.11	0.43	3.66
kk	0.00	0.00	0.00	0.00	0.00	0.04	0.04
kmr	-9.68	0.00	0.02	0.00	0.00	0.00	0.02
la	-0.06	0.52	0.19	1.58	0.35	0.28	2.36
la_ittb	-0.06	0.02	-0.09	1.48	0.05	0.45	1.89
la_proiel	-0.10	0.02	0.30	2.15	0.65	0.74	3.83
lv	0.20	0.02	0.63	0.21	0.25	0.03	1.12
nl	0.49	0.05	0.59	0.24	0.24	0.63	1.73
nl_lassysmall	0.38	0.05	0.34	0.34	-0.02	0.46	1.15
no_bokmaal	0.04	0.14	0.54	1.73	0.23	0.37	3.03
no_nynorsk	0.13	0.17	0.55	1.42	0.18	0.48	2.83
pl	-0.38	0.02	0.60	0.00	0.49	0.37	2.05
pt	-0.14	0.03	-0.51	0.06	0.00	0.35	-0.09
pt_br	-0.02	0.06	0.58	0.20	0.10	0.39	1.27
pt_pud	0.02	0.10	0.38	0.17	0.03	0.45	1.04
ro	0.91	0.16	1.20	1.68	0.60	0.57	4.15
ru	-0.62	0.16	-0.02	2.85	0.18	0.51	3.84
ru_pud	-0.71	0.00	0.00	3.84	0.36	0.50	5.05
ru_syntagrus	-0.65	0.10	0.04	2.38	0.18	0.55	3.43
sk	-0.75	0.10	0.49	1.12	0.09	0.83	2.67
sl	-0.09	0.00	0.73	2.19	0.26	0.79	4.19
sl_sst	0.11	0.01	1.21	1.30	0.40	0.14	3.62
sme	0.00	0.00	0.00	0.00	0.00	0.04	0.04
sv	1.07	0.11	1.38	2.12	-0.01	0.53	4.17
sv_lines	-0.36	0.18	1.47	2.32	0.22	0.33	4.71
sv_pud	-0.11	0.19	1.55	2.00	0.06	0.40	4.10
tr	0.28	0.00	0.43	0.00	0.26	-0.02	0.74
tr_pud	-0.12	0.00	0.23	0.00	0.07	0.05	0.31
ug	0.00	0.00	0.15	0.00	0.04	0.00	0.19
uk	-0.03	0.03	1.20	0.13	0.10	0.82	2.20
ur	0.03	-0.01	0.44	0.92	0.03	0.06	1.43
vi	-0.53	0.02	0.13	0.57	0.00	0.05	0.75
zh	-0.01	0.05	1.37	0.72	0.00	0.00	2.09

Table 6: Ablation shown by differences in UAS and LAS values.

5 Conclusion

For the CoNLL 2017 Shared Task on multilingual parsing from raw text, we were able to achieve a whole multilingual parser pipeline in a "semi-universal" manner exploiting minimum statistics from the training corpora with deterministic rules for part of speech tagging and label adjustment. Even with a simple and general model we achieved .43 labeled attachment scores on average and showed that the model we propose can be suitably applied to cross-lingual and universal scenarios.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *TACL* 4:431–444. http://aclweb.org/anthology/Q16-1031.

Y.J. Chu and T.H. Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica* 14:1396–1400.

J. Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards* 71(B):233–240.

Hiroshi Kanayama, Youngja Park, Yuta Tsuboi, and Dongmook Yi. 2014. Learning from a neighbor: Adapting a Japanese parser for Korean through feature transfer learning. *LT4CloseLang 2014* page 2. http://aclweb.org/anthology/W14-4202.

Héctor Martínez Alonso, Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Parsing universal dependencies without training. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 230–240.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 523–530. www.aclweb.org/anthology/H05-1066.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0 — CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-2184.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richard

Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Haji, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missila, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.0.

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the reproducibility of PAN's shared tasks: Plagiarism detection, author identification, and author profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*. Springer, Berlin Heidelberg New York, pages 268–299. https://doi.org/10.1007/978-3-319-11382-1_22.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association, Portorož, Slovenia.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. *CoRR* abs/1506.06158. http://arxiv.org/abs/1506.06158.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Hěctor Martínez Alonso, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.