How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection

Hiroshi Kanayama*, Ran Iwamoto†
*IBM Research - Tokyo
19-21 Nihonbashi Hakozaki-cho, Chuo-ku, Tokyo 103-8510 Japan
hkana@jp.ibm.com
†Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522 Japan
r.iwamoto@keio.jp

Abstract
This paper investigates clause-level sentiment detection in a multilingual scenario. Aiming at a high-precision, fine-grained, configurable, and non-biased system for practical use cases, we have designed a pipeline method that makes the most of syntactic structures based on Universal Dependencies, avoiding machine-learning approaches that may cause obstacles to our purposes. We achieved high precision in sentiment detection for 17 languages and identified the advantages of common syntactic structures as well as issues stemming from structural differences on Universal Dependencies. In addition to reusable tips for handling multilingual syntax, we provide a parallel benchmarking data set for further research.

Keywords: Universal Dependencies, sentiment analysis, multilinguality, parsing

1. Introduction
Sentiment analysis and opinion mining (Pang and Lee, 2008) along with their multilingualization (Korayem et al., 2016) have been studied for many years. In a typical use case for enterprises that continually seek information to improve their products or services while estimating future demands, it is very important to detect the individual utterances that specify positive or negative properties, beyond estimating the overall preference (typically, a number of stars) written in a review.

In this work, we pursue clause-level sentiment detection, which shares a similar motivation to aspect-based sentiment analysis (ABSA) (Pontiki et al., 2016). Our goal is to explore a general purpose system that doesn’t require data-specific features such as user and time information in Twitter data or Amazon reviews.

Most of the prior work has exploited machine learning for sentiment classification (Pang et al., 2002; Wang et al., 2012) or dictionary induction (Hamilton et al., 2016). However, it has been pointed out that statistical approaches to user-generated data (e.g. Twitter) may extract “iPod” as a positive keyword (Sath et al., 2012), and while this will improve the score in benchmarking datasets in a world where many people like iPods, we need to design a system free from such prior biases. Systems that can accurately detect positive and negative opinions are essential when it comes to solving other tasks, such as the recognition of sarcasm (Tungamahthiti et al., 2014) and social analysis to investigate race bias (Merullo et al., 2019).

To achieve a multilingual system that meets these requirements, we utilize a pipeline approach for clause-level sentiment detection. Our approach applies sentiment lexicon and syntactic rules to the output of dependency parser based on Universal Dependencies (UD) (Nivre et al., 2016). By capturing syntactic phenomena such as coordination and negation, we can accurately extract positive and negative polarities along with their corresponding predicate and targets by means of lexicon that can be manually configured or statistically expanded.

In this paper, rather than discussing a specific application, we focus on the role of the syntax layer. Specifically, we show how the dependency structures represented by UD and the dependency parsers contribute to multilingual sentiment detection. Through a series of experiments on 17 languages, we clarify the characteristics of UD structures and parsers, and demonstrate the advantages of multilingual SA. The main contributions of this paper are as follows.

1. Establish a methodology of multilingual semantic analysis on top of dependency structures by applying tree scanning, induced lexicon and valence shifters, and demonstrate that it can achieve high precision. (Section 4)

2. Evaluate the universality of Universal Dependencies technologically rather than linguistically by investigating the effects of language-universal and language-specific operations on the application. (Section 5)

3. Provide a multilingual resource for 19 languages generated from the parallel UD corpora to accelerate research on multilingual syntax. (Section 6)

2. Related Work
2.1. Universal Dependencies
Universal Dependencies (Nivre et al., 2016) (Nivre and others, 2019) is a worldwide project to provide a multilingual syntactic corpus. As of November 2019, 157 treebanks in 90 languages have been released. For all languages the syntax is represented by dependency trees with 17 PoS tags and 38 dependency labels commonly used for all languages, and each treebank can have language specific extensions. The resources and documentations are available online and incrementally updated. As a result

https://universaldependencies.org/
Two sentiment clauses (predicates and targets). The dependencies in bold lines from the root node are traversed to detect two sentiment clauses (predicates and targets).

2.2. Sentiment Analysis and Lexical Induction

For the development and evaluation of sentiment annotation on the level of phrase and clause rather than sentence or document, the Stanford Sentiment Treebank (Socher et al., 2013) is widely used as a dataset. Using this treebank, Verma et al. (2018) investigated popular sentiment annotators and noted their weakness in handling syntactic phenomena such as negation in subjects.

SentiWordNet (Psaltis and Sebastiani, 2006) and its extension (Baccianella et al., 2010) are widely used as a synset-level resource providing polarity with numerical degrees between negative one and one. Other lexicons based on semantic orientation have also been created (Esuli and Sebastiani, 2006) discussed in Osborne and Maxwell (2015). UD design is plausible for multilingual syntactic operation as shown in Section 3.2.

3. Clause-level SA

Our approach to clause-level sentiment analysis is aimed at fine-grained detection with high precision. This concept was originally discussed in a transfer-based sentiment extraction method analogous to translation (Kanayama et al., 2004). The main objective of clause-level SA is to detect polar clauses associated with a predicate and target. As an example, sentence (1) below conveys two polarities: (1a) a positive polarity regarding the hotel (which is loved) and (1b) a negative polarity about the waiters (who are not friendly).

(1) I love the hotel but she said none of the waiters were friendly.

(1a) + love (hotel)
(1b) − not friendly (waiter)

The rule-based sentiment analysis proposed by Vilares et al. (2017) shares a similar motivation to ours. They formalized a bottom-up operation to calculate the semantic orientation of each node on a syntactic tree of the Universal Treebank (McDonald et al., 2013) in a function-head style and tested their system on three languages. In this paper we use simpler rules without numerical values in a top-down manner of matching on content-head syntactic trees based on current Universal Dependencies in a content-head style, and cover more languages including diverse language families to clarify the effects of the multilingual syntax framework.

Figure 1: Dependency tree for sentence (1). The dependencies in bold lines from the root node are traversed to detect two sentiment clauses (predicates and targets).

Figure 2: Flow of clause-level sentiment detection.

The rule-based sentiment analysis proposed by Vilares et al. (2017) shares a similar motivation to ours. They formalized a bottom-up operation to calculate the semantic orientation of each node on a syntactic tree of the Universal Treebank (McDonald et al., 2013) in a function-head style and tested their system on three languages. In this paper we use simpler rules without numerical values in a top-down manner of matching on content-head syntactic trees based on current Universal Dependencies in a content-head style, and cover more languages including diverse language families to clarify the effects of the multilingual syntax framework.

3. Clause-level SA

Our approach to clause-level sentiment analysis is aimed at fine-grained detection with high precision. This concept was originally discussed in a transfer-based sentiment extraction method analogous to translation (Kanayama et al., 2004). The main objective of clause-level SA is to detect polar clauses associated with a predicate and target. As an example, sentence (1) below conveys two polarities: (1a) a positive polarity regarding the hotel (which is loved) and (1b) a negative polarity about the waiters (who are not friendly).

(1) I love the hotel but she said none of the waiters were friendly.

(1a) + love (hotel)
(1b) − not friendly (waiter)

We have build a system, as shown in Figure 2, to meet the requirements described in Section 1. The proposed system relies on dependency structures based on Universal Dependencies. This section describes the baseline implementation of the clause-level sentiment detector for English, as illustrated in Figure 2. First, the clauses that may convey sentiments
are detected in a top-down manner on dependency trees, and second, the polarities of the clauses are determined by using the sentiment lexicon and valence shifters.

### 3.1 Clause Detection

The main clause of a sentence is detected as the root node of the dependency tree, and its polarity is examined by matching with the lexicon described later. A single sentence may have multiple sentence clauses, so when the node has child nodes labeled `conjunction`, `parataxis` or `list`, those nodes are recursively scanned as potential sentiment clauses. When a node is a verb such as “say” or “think” that takes a `ccomp` (clausal complement) child, the child node is also examined. In example (1), two clauses, headed by “love” and “friendly” are detected (refer back to Figure 1). In addition to `conj`, a subordinate clause with an `advcl` label is examined if it has a marker such as “because”, “though” or “despite” labeled by `mark`, to cover (2). Otherwise subordinate clauses are not examined. Similarly to `ccomp`, an `xcomp` (open clausal complement) child of verbs such as “make” in (3) is subject to sentiment clause searches.

1. **[2]** `[+]` Because amenities were great, I was satisfied.
2. **[3]** `[+]` He made the travel excellent.

Sentences (1) and (2) include the positive adjective “beautiful”, but they do not form positive clauses.

1. I would go there if the bathroom is beautiful.
2. I want to stay in a beautiful room.

When we perform the clause detection in a top-down manner, “beautiful” in (1) and (2) can be excluded from the output due to the absence of a rule to examine a subordinate clause marked by “if” and an `xcomp` child of “want”. Note that manual annotation of a sentence or the output of statistical approaches may conclude that “beautiful”, but they do not form positive clauses.

### 3.2 Matching with Sentiment Lexicon

The sentiment lexicon consists of lexical entries associated with a lemma, a PoS tag, its polarity and the case frame. Table 1 shows some examples. Entry (a) is for the verb “love”, which is positive and takes a subject and a direct object; the target (which is positive) is its direct object. For most adjectives the target is in the subject, as in (b) “friendly”, but (c) “unhappy” specifies the target as “with”, which matches an `obl` child preceded by “with”, to detect “breakfast” as the target in (3). The lexicon has more expression power for disambiguation. Entry (d) for the adjective “high” is used only when the subject is “price”. Entry (b) can also match a noun phrase in `amod` relation as in (3), and the modified noun is the target.

### Table 1: Examples of lexical entries consisting of a lemma, part-of-speech, polarity, and a case frame.

| Entry | Word | Part-of-speech | Polarity | Case Frame |
|-------|------|----------------|----------|------------|
| (a)   | love | VERB           | +        | nsbj, obj  |
| (b)   | friendly | ADJ | +        | nsbj       |
| (c)   | unhappy | ADJ | -        | nsbj, with |
| (d)   | high   | ADJ           | -        | nsbj: "price" |
| (f)   | increase | VERB | =        | [nsbj]     |
| (g)   | increase | VERB | =        | nsbj.[obj] |
| (i)   | effectively | ADV | +        |             |

1. **[4]** The word in bold indicates the polar predicate and the underlined word is its target.

### 3.3 Valence Shifters

In addition to the inversion (as in (h) above), there are plenty of types of negation expression that reverse the polarities as studied extensively (Wiegand et al., 2011). The basic types of negation are direct negation of the verb and the noun in (13) and (15).

1. **[5]** Conventional function-head dependency treats “be” as the head of its complement “noise” in (1).
many of the corpora (including UD, in the parsing result. We use the list of negation clues “not”, “never”, “no”, etc. that modify the polar clauses with advmod or det: labels. When they match on the syntactic tree, the polarities assigned with the lexicon are reversed. Adverbs such as “seldom” and “scarcely” reverse the polarity as well, but “not only” should not be treated as negation even if “not” modifies the polar words. In addition to that, the negation in the subject and object as we have seen in (11) should be captured, using the negation pronouns (e.g. “nothing”, “nobody”).

3.4. Initial Evaluation

To determine how the syntax-based approach works for English, we evaluated polar clause extraction on the Stanford Sentiment Treebank (Socher et al., 2013) which provides sentiment degrees for any subtrees in a sentence, in addition to word/sentence-level sentiment. Our extraction system prioritizes precision over recall, so here, we limit our evaluation to the polar clauses detected by the system and do not evaluate recall.

A total of 2,210 sentences from the test portion of the Stanford Sentiment Treebank were processed by the system. When the system detects a polarity, the sentiment degree of the corresponding subtree in the treebank is examined, for the clause segmented by conj. parataxis, and list nodes (i.e. the units shown in the bottom of Figure 1), which contains the predicate and the target node. The detection is judged as correct if the system output is positive and the polarity score in the treebank is higher than 0.5, or negative and lower than 0.5; otherwise (including cases of 0.5) the result is judged as wrong. We used UDPipe (Straka and Straková, 2017) and StanfordNLP (Qi et al., 2018) as English UD-compliant parsers. Both of them were trained with version 2.4 of the UD_English-EWT corpus.

| parser     | correct | wrong | precision |
|------------|---------|-------|-----------|
| UDPipe     | 442     | 105   | 80.8%     |
| +lex       | 556     | 117   | 82.6%     |
| StanfordNLP| 558     | 120   | 82.3%     |
| +lex       | 650     | 131   | 83.2%     |

Table 2: Precision of polar clause detection in Stanford Sentiment Treebank (English movie data). Rows “+lex” show scores when the lexicon induced from the training data was added.

(14) [-] The hotel was not good.
(15) [+] It was no problem.

In some corpora in Universal Dependencies, a ‘Polarity=Neg’ feature is attached to the words for negation, but many of the corpora (including UD_English-EWT) do not support this feature. Therefore, we handle the negation phenomena as follows. We use the list of negation clues “not”, “never”, “no”, etc. that modify the polar clauses with advmod or det: labels. When they match on the syntactic tree, the polarities assigned with the lexicon are reversed. Adverbs such as “seldom” and “scarcely” reverse the polarity as well, but “not only” should not be treated as negation even if “not” modifies the polar words. In addition to that, the negation in the subject and object as we have seen in (11) should be captured, using the negation pronouns (e.g. “nothing”, “nobody”).

4. Multilingualization of SA

Based on the English implementation of clause-level SA described in Section 3, we broaden the language coverage relying on the common syntactic structures on UD. In this paper we handle 17 languages listed in Table 2 considering the availability of resources. The UD parsers used in this study are based on the UD corpora listed in Table 3.

4.1. Lexicon Transfer

The sentiment lexicon described in Section 3.2 needs to be prepared for each language. Since techniques of lexicon building are not the main focus of this paper, we applied simple ways to transfer our in-house English lexicon into other languages.

First, we converted the lexical entries of adjectives and verbs into English sentences with the templates shown in Table 2 (e.g., “We can love XYZ.” for entry (a) in Table 2). Those sentences are translated using the Watson Language Translator. The translated sentences were then parsed with the model of the target language. After replacing adjectives and verbs with their PoS tags, lemma sequences were compiled for each language, and frequent patterns (such as [er, sein, ADJ] in German and [nous, pouvoir, VERB, XYZ] in French) were extracted. When a translation matched with one of these common patterns, the lemma of the verb or the adjective was picked up as a lexical entry for the target language, with the same polarity as in the original English lexicon. To increase the coverage of the lexicon, we also used multilingual word embeddings created by aligning anchor word pairs in bilingual dictionaries and identical words (Conneau et al., 2017). We obtained five closest words to an English polar word in terms of cosine similarity in the embedding space of the target language. For each obtained word, its precision was evaluated and the highest score was selected.

Table 3: Sample templates of English sentences for lexicon translation.

| PoS | caseframe | generated sentence |
|-----|-----------|--------------------|
| ADJ | nsubj | “He is [adj].” |
| VERB | nsubj | “We can [verb].” |
| VERB | nsubj, obj | “We can [verb] XYZ.” |

Table 3: Sample templates of English sentences for lexicon translation.

4StanfordNLP and UDPipe achieved LAS (Labeled Attachment Score) of 86 and 78 on UD_English-EWT, respectively, when sentence boundaries were given. According to an investigation on parsing and SA (Gómez-Rodríguez et al., 2019), LAS=80 is considered good enough for SA.

5https://language-translator-demo.ng.bluemix.net/

6For Czech and Turkish, the polarity assigned to the lemma is reversed from the original one when the translated word has a “Polarity=Neg” feature in the parsing result.
was used. This demonstrates that two methods are comple-
mentary in terms of increasing the coverage.
For comparison, we also used Chen and Skiena (2014)’s
multilingual sentiment lexicon (“MSL”), which provides
lists of positive and negative words on surface forms. To
use it in our system, the lemma and PoS tags of the posi-
tive and negative words were filled by matching the lexical
entries of MSL and the form columns of the UD training
corpora. The numbers of obtained words are shown in the
right column of Table 3.

| language (en) | ours | MSL |
|--------------|------|-----|
| Arabic (ar)  | 499  | 1,727 |
| Czech (cs)   | 1,491 | 1,738 |
| German (de)  | 1,906 | 2,138 |
| Spanish (es) | 1,665 | 2,332 |
| Finnish (fi) | 1,101 | 1,390 |
| French (fr)  | 1,375 | 2,652 |
| Hebrew (he)  | 490  | 1,169 |
| Indonesian (id) | 641 | 1,121 |
| Italian (it) | 1,512 | 2,284 |
| Japanese (ja) | 385 | 225 |
| Korean (ko)  | 584  | 621 |
| Dutch (nl)   | 1,030 | 1,887 |
| Portuguese (pt) | 1,787 | 1,791 |
| Russian (ru) | 1,374 | 2,340 |
| Turkish (tr) | 286  | 653 |
| Chinese (zh) | 737  | 96  |

Table 4: Numbers of lexical entries per language. In ‘ours’
columns, ‘trans’ and ‘emb’ indicate the number of entries
obtained by translation and embeddings, respectively (‘-’
indicates the resource is not available). ‘union’ is the size
of the final lexicon, except for English which shows the size
of the original lexicon. ‘MSL’ column shows the numbers
of words obtained from Chen (2014)’s lexicons.

| language | PADT | ja | ISDT |
|----------|------|----|------|
| cs       | PDT  | ja | GSD  |
| de       | GSD  | ko | GSD  |
| en       | EWT  | nl | Alpino |
| es       | Ancora | pt | Bosque |
| fi       | TDT  | tr | IMST |
| fr       | GSD  | ru | SynTagRus |
| he       | HTB  | zh | GSD  |
| id       | GSD  |    | + GSDSimp |

Table 5: The names of UD corpora used in this study.

Syntactic Operations
As stated earlier, this paper examines the universality of UD
by applying syntactic operations to find appropriate
clauses and give the polarity after matching with the lex-
icon. Here, we focus on clause detection rules and valence
shifters, classifying them into language-universal, lexically
parameterized, and language-specific operations.

Clause detection The baseline of the clause detection can
be done in a language-universal manner: we just pick up
the clause of root and recursively follow its child nodes
labeled conj, parataxis, or list. Finding the clauses in
the child of ccomp and xcomp is a lexically parameter-
ized operation, with lists of head words for each language,
e.g., “think” and “make” for English and “creer” (believe)
and “parecer” (seem) for Spanish. These verbs can be
listed by searching for frequent words modified by ccomp
and xcomp in the corpora. Instances of language-specific
operations are described in Section 4.3.

Valence shifter The universal way to handle negation is
to rely on the “Polarity=-Neg” feature, which is available in
ar, cs, de, es, fr, he, id, pt, tr, and zh. As “not” in English,
11 languages (de, en, es, fr, he, id, it, nl, pt, ru, and zh)
have adverbs or particles for negation that modify the head
word using an advmod label. These are handled as lexical
parameters, the same as “no” in English with a det label.
Other languages require different ways to detect negation.
In Finnish and Japanese, auxiliary verbs for negation mod-
ify the head node with aux, and a negative copula (cop)
is used in Arabic. In Czech and Turkish, a verb or adjecti-
ve can be changed to its negative form while keeping its
lemma, thus “Polarity=Neg” is the only clue of negation. In
Korean, a negation form of a verb/adjective is represented
as a multi-word expression connected with flat.

Language-specific Issues
While the common syntactic structures of Universal Depen-
dencies are useful to design a multilingual system, we also
added language-specific operations, as their absence may
significantly reduce the performance, or even block all of
sentiment detection. The workarounds here help develop-
ners of multilingual downstream components that are based
on Universal Dependencies.

Arabic The lemma in the UD_Arabic-PADT corpus is
vocalized with Arabic tashkil marks, while normal sen-
tences are written without them. Both StanfordNLP and
UDPipe which were trained the corpus try to recover tashkil
marks with a lemmatizing accuracy of around 90%, but this
causes a mismatch between the input and the lexicon, so we

3For Japanese, manually generated lexicon was merged with
‘union’ for better testing of parsers and UD.
remove the tashkil marks when lemmata of the input words and lexical entries are compared.

**Czech** When a verb or adjective is negated with a prefix “ne”, “Polarity=Neg” is added to the feature in UD. However, sometimes the parser keeps “ne” in the lemma while the feature has “Polarity=Neg”, and this causes a wrong polarity (i.e., if “nеспокojený” (“unsatisfied”) is a lemma, the word shouldn’t have the negation feature), especially in the lexicon creation process. Our conservative workaround is to exclude words that have the negation feature and a lemma that still starts with “ne” in the lexicon transfer process, and to perform a similar operation during runtime.

**German** Adverbs and adjectives have same surface forms, so we handle ADV and ADJ interchangeably when matching the input with the lexicon so that we can detect polar adverbs with lexical entries for adjectives.

**Japanese** The conj label is never used in the UD Japanese corpora to avoid left-headed coordination structures that confuse the syntactic representations (Kanayama et al., 2018). To handle multiple clauses in a sentence, child nodes with an advcl label are examined for clause detection, with some exceptions for markers such as “て” (“if”).

**Korean** In the Korean UD corpora, the word unit is based on an eojeol (a phrasal unit split by whitespaces) and a lemma is expressed by the combination of all morphemes in the word connected with ‘+’ marks, which never matches the base form of the lexical entries. Thus the lemma form in a parsed result is converted into base form by picking the surface before the first ‘+’ mark and attaching a suffix “da”.

**Chinese** In this work we built a lexicon based on simplified Chinese characters, but the UD Chinese-GSD and -PUD corpora use traditional characters, so we need to switch the training model accordingly. UD Chinese-GSDSimp for simplified Chinese is available since UD version 2.5. A single-letter adjective preceded by “的情” (“very”) or “不” (“not”) tends to be regarded as a single word with the prefix, e.g., “很快” (“quick”). To increase recall, the prefixes are detached from the lemma, and the polarity is reversed when “不” is detached.

### 5. Evaluation

Unlike machine-learning methods with fixed training and test sets, it is not easy to fairly evaluate rule-based systems. To ensure transparency and a bias-free system while avoiding data overfitting, we roughly estimate the performance using existing datasets and focus on the relative comparison among languages, corpora, and types of syntactic operations, rather than comparing with other systems.

#### 5.1. Datasets and Metrics

To the best of our knowledge, there is currently no multi-lingual complete phrase-level sentiment annotation like that provided in the Stanford Sentiment Treebank, so we manage the evaluation of our system with sentence-level sentiment annotations. For Arabic, English, Spanish, French, Dutch, Russian, Turkish and Chinese, we use the dataset from the SemEval Workshop 2016 Task 5 for aspect-oriented sentiment analysis (Pontiki et al., 2016). XML data with aspect-level or sentence-level annotation is converted into a simple format: pairs consisting of a sentence and its binary polarity (positive or negative) without numerical degrees. In addition to polarities, our system outputs the sentiment targets, but we don’t evaluate them in this study because our notion of target is different from the aspect in those datasets.

To cover more languages, we added Amazon reviews used in a German shared task (Kuppenhofer et al., 2014), restaurant review data for Indonesian (Gojali and Khodra, 2016) and Czech (Steinberger et al., 2014), hotel review data for Italian (Balse et al., 2018), newswire data for Hebrew (Aniram et al., 2018), movie review tweets for Korean (based on the method by Maas et al. (2011)), book review data for Portuguese (Freitas et al., 2014) and opinions on mobile phones for Japanese (Hashimoto et al., 2011). All of these were converted into the common structures of the set of sentences with positive or negative flags. The statistics of the simplified data are shown in Table 6.

| language | genre | + | - | length |
|----------|-------|---|---|-------|
| ar       | hotel | 250 | 250 | 29.6  |
| cs       | restaurant | 250 | 250 | 16.5  |
| de       | cutlery | 297 | 62  | 13.7  |
| en       | restaurant | 250 | 250 | 14.7  |
| es       | restaurant | 250 | 250 | 15.4  |
| fr       | restaurant | 250 | 250 | 16.0  |
| he       | news | 250 | 250 | 14.2  |
| id       | restaurant | 250 | 250 | 10.7  |
| it       | hotel | 250 | 250 | 15.1  |
| ja       | mobile | 238 | 295 | 21.5  |
| ko       | movie | 250 | 247 | 9.4   |
| nl       | restaurant | 250 | 250 | 14.9  |
| pt       | book | 250 | 250 | 22.6  |
| ru       | restaurant | 250 | 250 | 17.3  |
| tr       | restaurant | 250 | 250 | 10.3  |
| zh       | mobile | 253 | 247 | 35.1  |

Table 6: Statistics of datasets for 16 languages used in this study. “+” and “−” are numbers of positive and negative sentences, respectively. “length” is the average number of words per sentence.

---

8This is not just a parser’s problem: the UD_Czech-PDT corpus has inconsistency in negation.
9Dutch has a similar syntax but such adverbs are tagged ADJ in UD corpora and thus no special care is needed.
10When aspect-level annotation is available, we picked up the sentences annotated with one or more consistent aspect-level polarities.
11Neutral sentences are discarded.
Table 7: Precision (prec), recall (rec), and $F_2$ score of sentiment detection tested on sentence-level datasets in 16 languages (%).

| lang | ours prec | ours rec | ours $F_2$ | MSL prec | MSL rec | MSL $F_2$ |
|------|-----------|----------|------------|----------|---------|-----------|
| ar   | 95.5      | 36.8     | 72.4       | 83.5     | 34.0    | 64.7      |
| cs   | 88.3      | 36.6     | 68.8       | 72.5     | 32.4    | 58.1      |
| de   | 94.2      | 46.8     | 78.3       | 85.6     | 56.5    | 77.6      |
| en   | 92.7      | 46.8     | 77.5       | 90.8     | 45.4    | 75.7      |
| es   | 90.2      | 36.6     | 69.8       | 74.0     | 38.4    | 62.4      |
| fr   | 90.0      | 43.6     | 74.2       | 76.1     | 53.6    | 70.2      |
| he   | 82.0      | 16.4     | 45.6       | 76.0     | 28.4    | 56.9      |
| id   | 92.7      | 33.8     | 68.7       | 83.5     | 33.0    | 63.9      |
| it   | 88.3      | 29.8     | 63.4       | 80.0     | 42.8    | 68.2      |
| ja   | 92.2      | 33.2     | 68.0       | 66.7     | 6.2     | 22.6      |
| ko   | 80.6      | 10.9     | 35.4       | 71.4     | 7.0     | 25.1      |
| nl   | 90.8      | 41.6     | 73.4       | 73.8     | 50.6    | 67.6      |
| pt   | 83.2      | 32.4     | 63.3       | 78.1     | 43.8    | 67.5      |
| ru   | 90.1      | 30.4     | 64.7       | 72.5     | 39.0    | 61.9      |
| tr   | 91.2      | 17.0     | 48.7       | 64.9     | 13.0    | 36.1      |
| zh   | 88.2      | 24.6     | 58.1       | 75.7     | 22.0    | 50.9      |

5.3. Comparison of UD Versions

To determine how the recent updates of Universal Dependencies corpora are helping multilingual operations, we applied our sentiment annotator to the 16 languages after parsing by three UDPipe models trained by UD versions 2.0, 2.2, and 2.4. Table 8 compares the three versions.

In Japanese, UD2.4 performed best because it has been updated to correctly assign advcl and acl. In Dutch, the score was significantly improved in UD2.2. This is due to improvement of parsing accuracy, and also to the updated attachment and labeling of adverbs including “niet” (‘not’) (advmood) in the corpora and parsers trained on them. UD2.0 corpora for Indonesian and Korean do not provide lemmata, and nor are the parsers trained on them, so the system could not detect any sentiment clauses. Despite the expectation for updates of UD corpora to improve this task incrementally, the scores were dropped in UD2.4 in some languages. For example, Russian parsed by the UD2.4 model showed significant changes in lemmatization that caused failures of matching with the lexicon. These findings should help facilitate further improvements of corpus annotation and parsing models.

5.4. Effects of Syntactic Operations

Here we examine how each syntactic operation described in Section 4.2 improved the sentiment detection, focusing on clause detection from subtrees (i.e. non-root nodes) and negation handling. Table 9 lists their ablation and stepwise improvements of scores.

Clausal complements with RM structures labeled as ‘not’ (‘niet’) (advmood) in the corpora and parsers trained on them. UD2.0 corpora for Indonesian and Korean do not provide lemmata, and nor are the parsers trained on them, so the system could not detect any sentiment clauses. Despite the expectation for updates of UD corpora to improve this task incrementally, the scores were dropped in UD2.4 in some languages. For example, Russian parsed by the UD2.4 model showed significant changes in lemmatization that caused failures of matching with the lexicon. These findings should help facilitate further improvements of corpus annotation and parsing models.

Forms. Approximately 20% of the errors were due to mismatch of polarities between sentence and clause, which are not real errors in practice. Remaining errors were caused by mishandling of syntactic phenomena, a lack in the domain lexicon, sarcasm, etc.

For all of the languages except Italian, Hebrew, and Portuguese, our translated lexicon (Section 4.1) contributed to higher scores than the MSL, especially in terms of precision, though MSL showed higher recall in eight languages. In Spanish, MSL had the negative entry ‘ir’ (‘go’), which frequently caused wrong detection. These results demonstrate the importance of maintaining a lexicon suitable for the system.
The "Polarity=Neg" feature is used in the '+U' operations that exploit dependency structures. This demonstrates the potential of UD from the viewpoints of applications that exploit dependency structures.

The right part of Table 8 shows the effects of negation handling, without which the precision is damaged. 'None' is the precision without handling any negation phenomena. The "Polarity=Neg" feature is used in the '+'UNV' situation. In Czech, Spanish, Portuguese, and Chinese, the errors were well reduced with this universal feature, but it did not change anything in seven of the languages. A negative contribution in French was caused by the typical negation expression "ne...pas", in which both "ne" and "pas" have a negation feature but they do not actually mean a double negation. Even with adding the words of advmod and det to each language, some negation phenomena were still not covered. In Japanese, Korean, and Arabic, all of the negation was expressed in language-specific ways and our operations recovered the precision.

Overall, the majority of phenomena are well covered by language-universal and lexically parameterized operations, with some exceptions (such as Japanese). This demonstrates the potential of UD from the viewpoints of applications that exploit dependency structures.

Table 8: Performance of sentiment detection with different versions of UD corpora. UDPipe’s parsing performance is shown as the labeled attachment Score (LAS).

| language | UD2.0 | UD2.2 | UD2.4 |
|----------|-------|-------|-------|
|          | LAS   | prec  | rec   | F2   | LAS   | prec  | rec   | F2   | LAS   | prec  | rec   | F2   |
| ar       | 64.3  | 88.4  | 15.8  | 46.1 | 65.1  | 85.3  | 16.8  | 47.0 | 66.6  | 83.7  | 17.0  | 46.9 |
| cs       | 82.3  | 85.4  | 30.4  | 62.7 | 82.8  | 86.6  | 29.8  | 62.7 | 82.9  | 86.4  | 31.8  | 64.3 |
| de       | 68.6  | 93.6  | 47.1  | 78.2 | 70.8  | 92.6  | 42.6  | 75.0 | 72.7  | 92.6  | 46.0  | 77.0 |
| en       | 76.5  | 92.6  | 43.8  | 75.7 | 77.1  | 92.5  | 44.0  | 75.0 | 76.4  | 90.9  | 43.4  | 74.6 |
| es       | 84.5  | 88.6  | 33.8  | 66.9 | 84.4  | 89.5  | 33.4  | 67.0 | 85.1  | 89.7  | 32.0  | 65.9 |
| fr       | 80.7  | 88.7  | 39.4  | 70.9 | 81.0  | 90.5  | 39.8  | 72.1 | 84.5  | 91.3  | 39.8  | 72.5 |
| he       | 57.9  | 82.1  | 14.0  | 41.6 | 57.9  | 82.1  | 14.0  | 41.6 | 58.3  | 82.3  | 13.2  | 40.2 |
| id       | 74.3  | –     | 0.0   | –    | 74.4  | 93.2  | 30.4  | 66.0 | 74.5  | 92.4  | 31.6  | 66.7 |
| it       | 86.1  | 85.7  | 25.6  | 58.3 | 86.3  | 89.1  | 25.0  | 58.9 | 86.7  | 85.5  | 25.2  | 57.8 |
| ja       | 75.5  | 92.5  | 29.5  | 64.8 | 72.6  | 90.1  | 28.9  | 63.3 | 76.2  | 92.1  | 32.6  | 67.5 |
| ko       | 60.5  | –     | 0.0   | –    | 61.4  | 83.3  | 9.1   | 31.7 | 61.4  | 83.7  | 8.2   | 29.5 |
| nl       | 69.6  | 90.6  | 25.8  | 60.3 | 77.6  | 90.6  | 39.4  | 71.9 | 77.6  | 89.4  | 39.8  | 71.6 |
| pt       | 82.5  | 81.9  | 29.2  | 60.2 | 82.2  | 79.9  | 27.0  | 57.4 | 82.7  | 87.8  | 27.8  | 57.6 |
| ru       | 87.3  | 88.9  | 26.4  | 60.3 | 84.6  | 88.9  | 30.0  | 63.8 | 85.0  | 87.3  | 27.6  | 60.9 |
| tr       | 55.8  | 92.6  | 15.2  | 45.9 | 54.0  | 94.7  | 14.6  | 45.2 | 55.1  | 94.8  | 14.8  | 45.6 |
| zh       | 57.7  | 76.4  | 8.6   | 29.6 | 57.7  | 81.2  | 7.8   | 28.2 | 58.7  | 84.4  | 7.6   | 27.9 |

Table 9: Ablation results of subtree search (for recall) and negation (for precision). ‘Δ+UNV’, ‘Δ+PRM’, and ‘Δ+SPC’ columns show the contribution to the metrics by language universal, lexically parameterized and language specific operations, respectively.
Adequate benchmarking data is still missing for some languages. To accelerate multilingual studies, we created a sentence-level sentiment corpus using parallel UD (PUD) corpora, each of which consists of 1,000 sentences. From the PUD corpora, parallel sentences with positive or negative polarities are extracted when our system detects a consistent polarity in four or more languages, assuming the sentiment polarity is shared in parallel sentences. We manually examined these sentences and filtered out wrong polarity assignments and politically biased decisions. A total of 106 polar sentences (48 positive and 58 negative) for 19 languages was obtained, including a language we didn’t evaluate in Section 5 (Finnish), and languages not covered in this study (Hindi, Thai, Swedish and Polish). We made this data available online (Kanayama, 2020).

Table 10 shows the results of the sentiment detection on the PUD data. Note that the data were created on the basis of our partial system outputs and thus it shows unfairly high precision, but recall is far from perfect in the languages with a relatively smaller lexicon (Arabic, Indonesian, and Turkish). In Chinese, the lexicon did not match well because the Chinese PUD corpus uses traditional characters. The issue in Korean has already been stated in the previous section. We can use these results for further improvement of systems and cross-lingual discussion of differences in syntax, by means of parallel visualization as exemplified in Figure 3.

Another advantage of using PUD corpora is that we can test the sentiment detection with the ‘gold’ dependency structures without caring about any parsing errors. The rightmost column in Table 10 shows the $F_2$ score on the gold syntax free from dependency errors, but unfortunately lemma is not provided in es, fr, id, ko, pt, and zh, and alphabetical lemmata are produced in Arabic, thus our system does not work at all for these languages. For other languages, the score is not always better than the results by StanfordNLP, due to the difference of annotations between main corpora and PUD. We suggest the unification of annotation policies in each language for further studies.

### 7. Conclusion

This paper has described multilingual sentiment detection that fully exploits the syntactic structures on Universal Dependencies. Thanks to UD’s common syntactic formalism, the system can cover many languages through the simple transfer of lexicon. Moreover, this work provided a methodology and reusable techniques for multilingual applications that do not require supervised data. Our analysis also revealed remaining issues with Universal Dependencies, such as word unit and lemmatization in Korean, and we provided parallel annotated data to accelerate future multilingual research.
8. Bibliographical References

Amram, A., Ben David, A., and Tsarfaty, R. (2018). Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern Hebrew. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2242–2252, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In LREC, volume 10, pages 2200–2204.

Basile, P., Croce, D., Basile, V., and Polignano, M. (2018). Overview of the EVALITA 2018 aspect-based sentiment analysis task (ABSITA). In Proceedings of the 6th evaluation campaign of Natural Language Processing and Speechtools for Italian (EVALITA18).

Berdichevskis, A., Coltink, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., Yan, C., Demberg, V., Lupyan, G., Rama, T., et al. (2018). Using universal dependencies in cross-linguistic complexity research. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 8–17.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 383–389.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. arXiv preprint arXiv:1710.04087.

Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic typology meets universal dependencies. In TLT, pages 63–75.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC, volume 6, pages 417–422. Citeseer.

Freitas, C., Motta, E., Milidiú, R. L., and Cesar, J. (2014). Sparkling Vampire... Io! Annotating opinions in a book review corpus. In In Sandra Aluisio & Stella E. O. Tagnin (eds.), New Language Technologies and Linguistic Research: A Two-Way Road, pages 128–146. Cambridge Scholars Publishing.

Gojali, S. and Khodra, M. L. (2016). Aspect based sentiment analysis for review rating prediction. In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), pages 1–6. IEEE.

Gómez-Rodríguez, C., Alonso-Alonso, I., and Vilares, D. (2019). How important is syntactic parsing accuracy? An empirical evaluation on rule-based sentiment analysis. Artificial Intelligence Review, 52(3):2081–2097.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2016, page 595.

Hashimoto, C., Kurohashi, S., Kawahara, D., Shinzato, K., and Nagata, M. (2011). Construction of a blog corpus with syntactic, anaphoric, and sentiment annotations. Journal of Natural Language Processing, 18(2):175–201.

Huang, J., Qiu, Q., and Church, K. (2019). Hubless neighbor search for bilingual lexicon induction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4072–4080, Florence, Italy, July.

Irvine, A. and Callison-Burch, C. (2017). A comprehensive analysis of bilingual lexicon induction. Computational Linguistics, 43(2):273–310.

Kanayama, H., Nasukawa, T., and Watanabe, H. (2004). Deeper sentiment analysis using machine translation technology. In Proceedings of the 20th International Conference on Computational Linguistics (COLING), pages 494–500, Geneva, Switzerland.

Kanayama, H., Han, N.-R., Asahara, M., Hwang, J. D., Miyao, Y., Choi, J. D., and Matsumoto, Y. (2018). Coordinate structures in Universal Dependencies for head-final languages. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 75–84.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. 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 2779–2795, November.

Korayem, M., Aljadda, K., and Crandall, D. (2016). Sentiment/subjectivity analysis survey for languages other than English. Social network analysis and mining, 6(1):75.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, pages 142–150.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The penn treebank.

McDonald, R., Nivre, J., Quirbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 92–97, Sofia, Bulgaria, August.

Merullo, J., Yeh, L., Handler, A., Grissom II, A., O’Connor, B., and Iyyer, M. (2019). Investigating sports commentator bias within a large corpus of American football broadcasts. 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 6356–6362, Hong Kong, China, November.
Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the second international conference on Knowledge capture*, pages 70–77, Sanibel, Florida.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia.

Osborne, T. and Maxwell, D. (2015). A historical overview of the status of function words in dependency grammar. In *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), pages 241–250, Uppsala, Sweden, August. Uppsala University, Uppsala, Sweden.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing* (EMNLP 2002), pages 79–86, Philadelphia, Pennsylvania.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October.

Ruppenhofer, J., Klinger, R., Struß, J. M., Sonntag, J., and Wiegand, M. (2014). IGGSA shared tasks on German sentiment analysis (GESTALT). In Gertrud Faabøt et al., editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 164–173, Hildesheim, Germany, October. Universität Heidelberg.

Saif, H., He, Y., and Alani, H. (2012). Alleviating data sparsity for Twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS.org).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Steinberger, J., Brychcin, T., and Konkol, M. (2014). Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland, June. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tunghamthiti, P., Shirai, K., and Mohd. M. (2014). Recognition of sarcasms in tweets based on concept level sentiment analysis and supervised learning approaches. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 404–413, Phuket, Thailand, December. Department of Linguistics, Chulalongkorn University.

Verma, R., Kim, S., and Walter, D. (2018). Syntactical analysis of the weaknesses of sentiment analyzers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1122–1127.

Vilares, D., Gómez-Rodríguez, C., and Alonso, M. A. (2017). Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, Jeju Island, Korea, July.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., and Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68.

Zeman, D., Ginter, F., Hajic, J., Nivre, J., Popel, M., and Straka, M. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium.

Zhao, M. and Schütze, H. (2019). A multilingual BPE embedding space for universal sentiment lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3506–3517, Florence, Italy, July.

9. Language Resource References

Hiroshi Kanayama. (2020). *Parallel Sentiment*. ELRA, ISLRN (awaiting).

Nivre, Joakim and others. (2019). *Universal Dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, ISLRN 586-682-285-530-1.