Automatic Morpheme Segmentation and Labeling in Universal Dependencies resources

Miikka Silfverberg and **Mans Hulden**
Department of Linguistics
University of Colorado
{miikka.silfverberg,mans.hulden}@colorado.edu

Abstract

Newer incarnations of the Universal Dependencies (UD) resources feature rich morphological annotation on the word-token level as regards tense, mood, aspect, case, gender, and other grammatical information. This information, however, is not aligned to any part of the word forms in the data. In this work, we present an algorithm for inferring this latent alignment between morphosyntactic labels and substrings of word forms. We evaluate the method on three languages where we have manually labeled part of the Universal Dependencies data—Finnish, Swedish, and Spanish—and show that the method is robust enough to use for automatic discovery, segmentation, and labeling of allomorphs in the data sets. The model allows us to provide a more detailed morphosyntactic labeling and segmentation of the UD data.

1 Introduction

Recent versions of Universal Dependencies (UD) (Nivre et al., 2017) provide not only part-of-speech labeling, but also universal lexical and inflectional features on most word forms. Table 1 illustrates a few example words from the three experiment languages used in this paper.[1]

A noteworthy aspect of this layout of the data is that it provides for an interesting inference problem in the realm of weakly supervised learning of inflectional morphology.[2] First, we note that the feature-value pairs in the annotation correspond mostly to individual allomorphs in the surface form of the word. For example, in the Spanish word **asignados** (Table 1), a standard analysis would be that the **asign-** part corresponds to the stem, the **-ad-** corresponds to `VerbForm=Part` and `Tense=Past`, the **-o-** to `Gender=Masc` and the **-s** to `Number=Plur`. The inference problem is then: given many annotated word forms with morphosyntactic features which are not matched to any substrings in the word, find a globally satisfactory segmentation of all word forms and associate the morphosyntactic labels in each word with these segmented substrings.

2 Related Work

Morphological segmentation, particularly in unsupervised scenarios, is a standard problem in NLP, and has been explored in numerous works (Goldsmith (2001), Creutz and Lagus (2005), Poon et al. (2009), Dreyer and Eisner (2011) inter alia). We recommend Ruokolainen et al. (2016) for an overview. Likewise, semi-supervised, or minimally supervised models—where the supervision usually implies access to some small number of segmented words—have also been widely investigated (Dasgupta and Ng, 2007; Kohonen et al., 2010; Grönroos et al., 2014; Sirts and Goldwater, 2013). Many approaches also take advantage of a semantic signal, or a proxy for semantic similarity between words such as Latent Semantic Analysis (Schone and Jurafsky, 2000) or its more modern counterpart, word embeddings (Soricut and Och, 2015). The specific formulation of an inference problem like the one presented in this paper has to our knowledge not been directly addressed previously, probably due to the necessity of annotated resource schemas such as those present in UD 2.0. A related problem, dealt with in Cotterell et al. (2016b) and Kann et al. (2016), concerns simultaneous segmentation and canonicalization—

[1] We have deviated slightly from the original annotation, incorporating the lemma as a feature for each word, the need for which will be explained in the technical portion of the paper.

[2] A similar annotation is provided in the SIGMORPHON shared task (Cotterell et al., 2016a) data set, although without implicit token frequency information since the data comes in the form of inflection examples mostly from Wiktionary.

Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 140–145,
Gothenburg, Sweden, 22 May 2017.

Finnish	`jäällä`	`Noun|Lemma=jää|Case=Ade|Number=Sing`
Spanish	`asignados`	`Verb|Lemma=asignar|Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part`
Swedish	`innebär`	`Verb|Lemma=innebära|Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Act`

Table 1: Examples of the modified UD annotations used for inference of segmentation and labeling.

a task where allomorphs are both segmented and rendered as a single canonical form, e.g. **communism** ↦ **commune ism**. This task was addressed in an entirely supervised scenario, however, and so the results are not directly comparable.

3 The Segmentation and Labeling Problem

As implied above, the current labeling of the UD data provides significant constraints and a supervision signal that can guide us in the inference process. One strong linguistically informed bias is that labels, i.e. abstractions of morphemes such as `Number=Sing`, `Gender=Masc`, should be assigned to substrings in such a way as to co-occur only with a small number of distinct strings throughout the data. This corresponds to the idea that each morpheme be realized as a limited number of distinct allomorphs. For example, the English pluralizer morpheme by and large occurs as only three allomorphs, **-s,-es,** ∅. Another intuition is the inverse of the previous one: that each allomorph only co-occur with a limited number of labels. For example, the **-s** allomorph in English serves mainly two distinct functions: a pluralizer and the third person present tense marker. We expect rampant ambiguity not to be present in the morphology of a language. On the whole, since most labels are seen a large number of times, we can develop a model that leverages this information to favor correspondences that are systematic in the data. Figure 1 illustrates a linguistically sound correspondence over several word forms that involve two stems in Spanish.

The intuition behind our model is that we'd like to find a segmentation of all words in the data into constituent allomorphs, and provide a label for each allomorph that fulfills the properties above. To perform this, we take advantage of the fact that we already know which morphemes (feature-value pairs) are present in each word (although some of these labels will correspond to null allomorphs).

In general, we want to explore the space of all possible segmentations and labelings in the data and find one that optimizes some objective function $\mathcal{C}$, based on the above observations. A given proposal segmentation S and labeling F of the data gives us a joint distribution $P_{S,F}$ over pairs of substrings $s \in \Sigma^*$ and labels $l \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of labels (feature-value pairs) used in the data. We can formalize a cost function $\mathcal{C}(S, F)$ based on the distribution $\mathrm{P}(S, F)$. This cost function could take many linguistically motivated specific forms: simply minimizing the total number of resulting distinct allomorphs in the data, minimizing the joint entropy of the labels and the allomorphs, maximizing the mutual information of the allomorphs and the labels, etc. Below, we use a specific cost function that maximizes a measure of symmetric conditional probability between segments and labels.

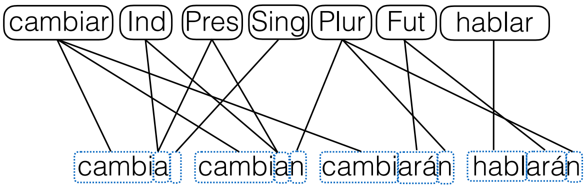

Figure 1: Morphosyntactic features are assigned corresponding substrings where re-use of the same label-substring correspondences is encouraged by the model. Note that some labels (such as `Sing` here) can be assigned to empty substrings.

4 Model

4.1 Definitions

Let $\mathcal{D} = \{(x_1, y_1), ..., (x_k, y_k)\}$ be a collection of word forms x_i and sets of associated morphological features y_i, for example

`dogs {lemma=dog,num=plural}`

As explained in Section 3, we learn a segmentation $S = \{s^1, ..., s^k\}$ of words in $\mathcal{D}$, where each $s^i = (s^i_i ... s^i_n)$ is a segmentation of word x_i into substrings, and a set of feature assignments $F = \{\mathrm{f}_i : y_i \to s^i | 1 \leq i \leq k\}$ of morphological features in y_i onto substrings in s^i.

Because of the existence of unmarked morphological features, such as singular number of nouns in English, we have to allow assignment of morphological features to a zero morpheme. We accomplish this by adding an empty substring to each segmentation.

Each segmentation and label assignment of the data set $\mathcal{D}$ defines joint counts $c(s, f)$ of substrings s and morphological features f as in Equation 1.

$$c(s, f) = \|\{s^i_j | s^i_j = s \text{ and } f_i(f) = s^i_j\}\| \quad (1)$$

Using $c(s, f)$ we express the probability of the co-occurrence of a feature and substring in Equation 2.

$$\mathrm{P}(s, f) \propto c(s, f) + \alpha B(s, f) \quad (2)$$

The function B in Equation 2 expresses a prior belief about the joint counts of segments and labels, and hyper-parameter α controls the weight of the prior information (Goldwater and Griffiths, 2007). A large α will result in $\mathrm{P}(s, f)$ which very closely reflects the prior belief while a smaller α lets P adapt more closely to the current segmentation and label assignment. We set α to 0.1 in all experiments.

We use the joint distribution of substrings and labels in the unsegmented data set $\mathcal{D}$ as prior information. Thus $B(s, f) = \#(s, f)/\#(f)$, where $\#(s, f)$ is the count of substrings s in words with morphological feature f and $\#(f)$ is the count of feature f in $\mathcal{D}$.

For lemma features, for example `lemma=dog`, we add an additional factor to the co-occurrence probability $\mathrm{P}(s, f)$ as shown in Equation 3. The quantity $\mathrm{d}(s, f)$ represents the edit distance of the substring s and the lemma corresponding to f. For example, d(**do**,`lemma=dog`) = 1. This allows us to model the fact that the stem and lemma of a word form often share a long common substring.

$$\mathrm{P}(s, f) \propto (c(s, f) + \alpha B(s, f)) \cdot 2^{-\mathrm{d}(s,f)} \quad (3)$$

4.2 Objective Function

Our objective function is the *symmetric conditional probability* over segments s and morphological features f defined by Equation 4

$$\mathcal{C}(S, F) = \prod_{s \in \Sigma^*, f \in \mathcal{Y}} \mathrm{P}(s|f)\mathrm{P}(f|s) \quad (4)$$

Symmetric conditional probability was introduced by da Silva et al. (1999) for multi-word expression extraction. The measure is intuitively appealing for our purposes since it is maximized when each morphological feature is associated with exactly one allomorph, and this allomorph, in turn, only occurs with the specific morphological feature.[3]

4.3 Inference

The space of possible segmentations and label assignments to each allomorph segment is very large except for toy data sets. Therefore, an exact solution to the optimization problem presented in Section 4.2 is infeasible. Instead, we use Gibbs sampling to explore the space of possible segmentations S and feature assignments A of our data set $\mathcal{D}$ with the intent of finding the segmentation S_{max} and assignment A_{max} which maximize the symmetric conditional probability of segments and features.

Gibbs sampling in this context proceeds by sampling a new segmentation S' and assignment A' from the current segmentation S and assignment A, and then either rejecting the old segmentation and assignment in favor of the new one with probability $(\mathcal{C}(A', S')/\mathcal{C}(A, S))^{\beta}$, or keeping the old segmentation and assignment. We set the hyper-parameter β to 2 in all experiments and run the Gibbs sampler on the data set $\mathcal{D}$ until the value of the objective function $\mathcal{C}$ has converged.

A new segmentation S' and label assignment A' can be sampled from an existing segmentation S and assignment A in two steps. First, randomly choose a word x_i from the data set. Using its current segmentation s^i in S, form the set of new segmentation candidates C by (1) joining two segments in s^i, (2) splitting one of the segments in s^i, or (3) moving a segment boundary in s^i one step to the left or right. The set C is illustrated in Figure 2.[4] Then randomly sample a new segmentation c from C.

Next, assign the labels in y_i to the segments of c in the following way. Iteratively, choose the substring $s \in c$ and feature $f \in y_i$ of maximal symmetric conditional probability $P(s|f)P(f|s)$, provided that no features have yet been assigned to s, and f has not been assigned to a substring. When each substring in c has been assigned exactly one label, assign remaining labels to substrings in c which maximize the symmetric conditional probability.

[3] This is, of course, not true in general because morphemes often have more than one allomorph. Nevertheless, the number of allomorphs is small for most stems and affixes.

[4] We assume that every non-empty segment has a corresponding morphological feature. Therefore, we filter out segmentations where the number of segments exceeds the number of morphological features y_i for the given word x_i.

(a)

	Finnish	Spanish	Swedish
Recall	87.43	84.38	88.71
Precision	94.63	88.63	94.01
F_1-score	**90.89**	**86.45**	**91.28**
Morfessor baseline			
Recall	80.65	81.32	90.82
Precision	76.92	73.64	75.58
F_1-score	78.74	77.29	82.50

(b)

	Finnish	Spanish	Swedish
Recall	62.79	50.10	55.87
Precision	71.06	54.22	61.82
F_1-score	**66.67**	**52.08**	**58.69**
Morfessor baseline			
Recall	30.51	25.93	44.13
Precision	28.45	22.24	32.92
F_1-score	29.45	23.94	37.71

(c)

	Finnish	Spanish	Swedish
Recall	80.07	73.49	88.26
Precision	90.62	79.54	97.66
F_1-score	**85.02**	**76.39**	**92.73**
Morfessor baseline			
Recall	74.96	48.34	83.10
Precision	69.90	41.47	62.00
F_1-score	72.34	44.64	71.01

Table 2: Results for (a) morpheme boundaries; (b) unlabeled morphemes; (c) labeled morphemes.

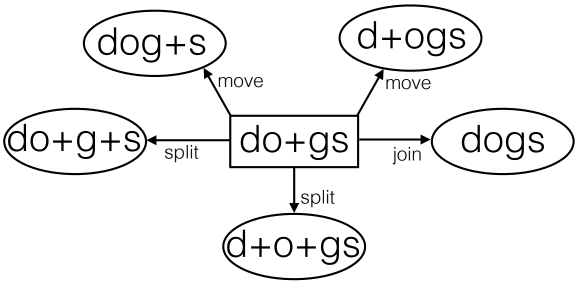

Figure 2: The set of new segmentation candidates for word *dogs* given the old segmentation *do+gs*. Each of the new segmentations is equally probable.

5 Experiments

We conduct experiments by running Gibbs sampling on words and morphological labels in the combined training and test data (without manual segmentations and label assignments). We then compare the segmentations and label assignments, discovered by the system, with the manually prepared annotations in the test data.

5.1 Baseline

As a baseline, we use the Morfessor system (Creutz and Lagus, 2005) for unsupervised segmentation.[5] We then assign labels to substrings as explained in Section 4.3. However, as we cannot control the number of segments given by Morfessor, we may end up with substrings to which we cannot assign morphological features. This happens in the case where the number of substrings given by Morfessor exceeds the number of morphological features for the word.

5.2 Data and Evaluation

We use three treebanks from the Universal Dependency v1.4 resource for experiments: `UD-Finnish`, `UD-Spanish` and `UD-Swedish`. We use the first 10,000 word forms from the training sets of each treebank for training (these contain 5,892 unique word forms for Finnish, 3,624 unique word form for Swedish and 4,092 unique word forms for Spanish) and the first 300 words from the test sets of each treebank for testing (these contain 253 unique word forms for Finnish, 172 unique word forms for Swedish and 278 unique word forms for Spanish). Punctuation and numbers were excluded from the training and test sets.

We remove a number of UD labels which do not express morphological categories, for example `style=arch` and `abbr=yes`.[6]

The test sets were manually segmented and morphological features were manually assigned to the segments by competent language speakers. The average number of morphemes per word in the test sets are 1.9 for Finnish, 1.7 for Spanish and 1.4 for Swedish, respectively.

We evaluate our system with regard to recall, precision and F_1-score for (1) morpheme boundaries including word boundaries, (2) unlabeled morphemes, and (3) labeled morphemes. In the case of labeled morphemes, a single substring can be counted multiple times if it has been assigned multiple morphological features. That is, even when the system fails to predict some of the morphological features correctly for a given substring, it will still receive a score for the features it did manage to predict correctly.

5.3 Results

Results are shown in Figures 2 (a), (b), and (c). The advantage given by leveraging the weak labeling in UD is visible in that the proposed system clearly outperforms the unsupervised Morfessor baseline for all languages.

Results for labeled morphemes are substantially

[5] We use revision 4219fbcc27ee0f5e3a4dca8de9f7ffc7a5bfe5e0 of `https://github.com/aalto-speech/morfessor` and default settings for all hyperparameters.

[6] Our data sets and code are publicly available at `https://github.com/mpsilfve/ud-segmenter`.

better than for unlabeled morphemes because the same substring can be scored as correct multiple times if it is associated to several morphological features. Moreover, the F_1-score for labeled morphemes is computed over both non-empty and empty substrings because morphological features can be realized as a zero morpheme. In contrast, the unlabeled morpheme F_1-score only considers non-empty substrings—i.e. the unlabeled segmentation is not rewarded for declaring empty allomorphs.

Overall, our system performs well on Finnish and Swedish but performance is markedly worse on Spanish—although an error analysis reveals that many of the incorrect segmentations in Spanish are linguistically defensible.

6 Discussion & Future Work

The system is immediately deployable for all UD languages and provides a segmentation and labeling of allomorphs, which may be useful for other downstream tasks. While the segmentation is not linguistically perfect, it is consistent. We also note that in many cases it is not linguistically clear-cut where morpheme boundaries should be drawn. An illustrative example is provided by Spanish verb forms where the infinitive, future, and conditional forms always contain an **-ar**, **-er**, or **-ir** substring, e.g. **hablar**, **comer**, **vivir**. Traditionally, the verb stem itself is not assumed to include these since, for example, subjunctive and some preterite forms surface without the vowel or the **r**: **hablé**, **comía**, **viva**. From an information-theoretic point of view, it is unclear which stem shape is an appropriate linguistic choice to declare. This is due to the fact that most witnessed forms in the data retain at least the vowel because present indicative forms are quite frequent, e.g. **hablan** (3P-PL) or **hablamos** (1P-PL), etc. Indeed, our algorithm chooses to include the vowel, probably because of the overwhelming frequency of present tense forms.

Our algorithm generally performs quite well on Finnish, however, there are a number of problematic morphological features which cause segmentation errors. For example, plural number for nouns, adjectives and pronouns is a source of errors. In Finnish, plural number in nouns and adjectives is realized by three different affixes **-i-**, **-j-** and **-t**. Pronouns are also marked for number in the UD data set but these affixes are not present in pronouns. Instead, plural number is realized as the zero morpheme in our gold standard segmentation. This means that there is a large number of different realizations for plural number, which may explain the fact that our system quite often incorrectly assigns plural number to the zero morpheme. Another problem is caused by illative case which is realized as **-Vn** or **-hVn** where **V** refers to the last vowel of the preceding word stem. As in the case of plural number, this leads to a large amount of different realizations for illative case. All of these, nevertheless, share the final suffix **-n**. Therefore, our system often prefers to drop all non-final characters and incorrectly marks illative case as **-n**.

The most frequent error in the Swedish data set is that the definite noun markers (**-en**, **-et**, **-n**) and adjective markers (**-a**) are assigned to the zero morpheme. This may be related to the fact that pronouns, which are quite common in the data set, are marked for definiteness but do not always carry the same affixes as nouns. For example, the Swedish pronoun **dessa** is definite but carries none of the definiteness markers for nouns. This can most likely be addressed by invoking separate models per part of speech so that the model is not confused by similar suffixes occurring with entirely different tags.

The majority of segmentation errors seem stem from the tendency of the SCP scoring to strongly prefer one-to-one correspondences between allomorphs and morphological features. Situations where a morphological feature can be realized by a large number of different allomorphs present problems. At the present time, solving these problems remains future work. To this end, we plan to experiment with different cost functions as the SCP appears to perform best on agglutinative languages where the one-to-one assumption holds stronger than for fusional languages. Likewise, root-and-pattern morphologies, such as found in the Semitic languages, have not been considered here since this would require permitting that allomorphs be discontinuous in a word form. Extending the model to handle such phenomena is straightforward, but requires associating labels with subsequences instead of substrings, which in turn greatly enlarges the search space, and requires efficiency improvements in the sampler to be able to handle large data sets where discontinuous morphemes are present.

References

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016a. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany, August. Association for Computational Linguistics.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016b. A joint model of orthography and morphological segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, San Diego, California, June. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Helsinki University of Technology.

Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Progress in Artificial Intelligence: 9th Portuguese Conference on Artificial Intelligence, EPIA '99 Évora, Portugal, September 21–24, 1999 Proceedings*, pages 113–132. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 155–163, Rochester, New York, April. Association for Computational Linguistics.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of EMNLP 2011*, pages 616–627, Edinburgh. Association for Computational Linguistics.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic, June. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1177–1185, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967, Austin, Texas, November. Association for Computational Linguistics.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, pages 78–86. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, et al. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217. Association for Computational Linguistics.

Teemu Ruokolainen, Oskar Kohonen, Kairit Sirts, Stig-Arne Grönroos, Mikko Kurimo, and Sami Virpioja. 2016. A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, 42(1):91–120.

Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, pages 67–72. Association for Computational Linguistics.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado, May–June. Association for Computational Linguistics.