Weakly supervised learning of allomorphy

Miikka Silfverberg and **Mans Hulden**
Department of Linguistics
University of Colorado
first.last@colorado.edu

Abstract

Most NLP resources that offer annotations at the word segment level provide morphological annotation that includes features indicating tense, aspect, modality, gender, case, and other inflectional information. Such information is rarely aligned to the relevant parts of the words—i.e. the allomorphs, as such annotation would be very costly. These unaligned weak labelings are commonly provided by annotated NLP corpora such as treebanks in various languages. Although they lack alignment information, the presence/absence of labels at the word level is also consistent with the amount of supervision assumed to be provided to L1 and L2 learners. In this paper, we explore several methods to learn this latent alignment between parts of word forms and the grammatical information provided. All the methods under investigation favor hypotheses regarding allomorphs of morphemes that re-use a small inventory, i.e. implicitly minimize the number of allomorphs that a morpheme can be realized as. We show that the provided information offers a significant advantage for both word segmentation and the learning of allomorphy.

1 Introduction

Many NLP resources provide weakly labeled morphological resources in data sets that are primarily annotated for higher-level constructs besides morphology. Most treebanks, for example, include some morphological annotation on the word level of varying granularity. The Penn treebank (Marcus et al., 1993) uses a limited label set of 45, while the Universal Dependencies (UD) (Nivre et al., 2017) project annotates word forms with a much larger set of morphological feature-value pairs. Noteworthy is that such annotation is not in any way aligned with the substrings in the word forms themselves: if the Finnish word **kaatuisi** 'would fall down' is annotated as `kaatua,V,Cond,Pres,3,Sg`, there is no indication that **kaatu** corresponds to the stem, **isi** to `Cond`, and that `V`, `3` and `Sg` are realized as zero allomorphs.

In essence, such labeled resources provide an inference problem in the realm of inflectional morphology in that one can exploit statistical regularities in the data to perform a morpheme segmentation and labeling of the data. A linguistically informed observation based on a simple assumption of systematic regularity between form and meaning is that it is very unlikely that a single morpheme such as the Finnish conditional be realized in more than a handful of different allomorphs. Conversely, it is unlikely that a part of a word, such as the affix **isi** carry many disparate meanings, i.e. be associated with a large array of different labels. Still, morphemes are often realized by more than one allomorph although the number of allomorphs is typically small. Consider for example English plural number which is realized by different allomorphs in the forms **dogs**, **churches**, **oxen** and **children**. From a data-driven perspective, the inference problem thus becomes to find a globally good allomorph segmentation and labeling of all word forms given in a large resource of inflected word forms.

Besides NLP applications, this type of input and the related inference problem is consistent with the assumptions of relevant inputs witnessed in L1 acquisition—a combination of stems and other affixes where the learner knows from the environment some semantic signal from the immediate discourse, e.g. plurality, tense, etc. Children tend

to show the ability to analyze affixes before they can use them productively. For example, three-year-olds have been shown to be able to associate agentive meaning to an **-er** morpheme in English, but can only produce the suffix later (Clark and Hecht, 1982; Clark and Berman, 1984).

In this paper we explore and evaluate several methods for automatically segmenting and labeling each allomorph present in resources that are labeled with morphosyntactic features at the word level. This means that our training data consists of plain unsegmented word forms (for example **kaatuisi**) and morphological feature sets (for example {V, Cond, Pres, 3, Sg}). The result is a morphologically segmented corpus where each morphological segment is associated with at least one morphological feature as shown in Figure 1. In order to account for fusional morphology, we allow one segment to be associated with multiple morphological features.

We treat the problem of joint segmentation and feature assignment as a search problem in the space of all possible segmentations and labelings of each word form in a (weakly) annotated corpus. The crucial constraint provided by the weak labeling is that not all labels can be present in a word form—the set of labels present for each inflected word must be restricted to those given by the resource. To our knowledge, this weakly supervised task has not previously been explored although joint segmentation and labeling has been explored in a fully supervised setting by Cotterell et al. (2015).

To solve the problem, we explore global metrics that indirectly favor re-use of allomorphs according to the intuition given above. We formalize a generic objective function that scores the goodness of segmentations and labeling globally in a corpus. The scoring portion of this objective function is tested with several metrics: symmetric conditional probability, which favors that allomorphs be good predictors of labels and vice versa, a perceptron learner that weights allomorph-label association, a Rescorla-Wagner model based on classical conditioning that also learns such association weights, and a model of Kullback-Leibler divergence that favors that labels and allomorphs have similar distributions throughout a data set.[1] We also compare the performance of the various methods to a baseline unsupervised model, Morfessor[2], augmented with the capacity to also provide labels of allomorphs in addition to segmenting.

2 Related Work

In the realm of natural language processing, morphological segmentation is a well-researched and established problem (Goldsmith (2001), Creutz and Lagus (2005), Poon et al. (2009), Dreyer and Eisner (2011), Ruokolainen et al. (2016)). While most approaches to pure segmentation are unsupervised, semi-supervised work usually assumes the availability of a limited number of gold segmentations (Dasgupta and Ng, 2007; Kohonen et al., 2010; Grönroos et al., 2014; Sirts and Goldwater, 2013). Using vector space representations of words to produce a weak labeling that identifies related forms has also been investigated (Schone and Jurafsky, 2000; Soricut and Och, 2015). Kann et al. (2016) perform unsupervised *canonicalization* of allomorphs, transforming words such as **having** to **have ing**, a task which is somewhat related to the problem addressed in this paper. Many methods that tackle specific morphology-related NLP tasks implicitly learn some model of allomorphy. This includes semi-supervised vocabulary expansion (Faruqui et al., 2016), and morphological inflection from examples (Cotterell et al., 2016a).

To our knowledge, the weakly supervised learning problem addressed in this paper has not been considered in the literature. Cotterell et al. (2015) present a closely related task. They investigate *labeled morphological segmentation*, that is, simultaneous segmentation and labeling of segments with morphological features. The crucial difference between our work and the work by Cotterell et al. (2015) is that our models are learned in a weakly supervised manner from plain word forms and sets of morphological features. In contrast, Cotterell et al. (2015) learn segmentation models in a fully supervised manner from data where each word is morphologically segmented and the segments are annotated with morphological features.

In the cognitive literature on L1 and L2 learning, statistical learning -based approaches that attempt to explain language learning through observations about statistical regularities have explored the extent to which relatively simple generalizations based on co-occurrence observations and

[1] Our code is freely available at https://github.com/mpsilfve/learn-allomorphs

[2] http://www.cis.hut.fi/projects/morpho/morfessor2.shtml

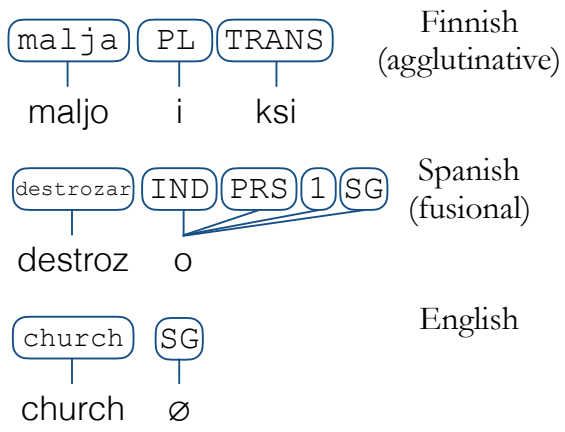

Figure 1: Morphological feature alignments in agglutinative and fusional languages; in the Finnish word (**malja** 'cup') each allomorph has a single feature while in the Spanish word (**destrozar** 'destroy') several features are associated with a single allomorph. In the English example, a zero allomorph is declared to which the feature SG is aligned.

statistical generalizations can be used to model learning of various levels of structure in natural language, including discovery of words (Saffran et al., 1996), grammatical categories (Reeder et al., 2013), and syntactic structure (Newport, 2016).

3 Methods

Given a corpus of word forms and associated morphological features, we want to find the correct segmentation for a word such as **barked** into *segments* which correspond to morphemes **bark**, **ed**, ∅, and the correct assignment of *morphological features* in a feature set {bark,V,Past} onto the segments. In this case: **bark**/bark, **ed**/Past and ∅/V. Note that we treat the lemma as a morphological feature.

Zero morphs (∅) are required because several languages have morphological features which are not visible in the word form, for example singular number of English nouns. In our gold standard segmented test data, we align word class markers such as N and V with a zero morph because they do not correspond to any substring of the word form. This decision is somewhat arbitrary. Other options include aligning them with the word stem and simply removing them from the corpus. In some cases, as in the case of English adverbs with suffix **-ly**, one could also consider aligning the word class marker with an affix.

We propose to accomplish segmentation and feature assignment by learning a real-valued scoring function $\Theta : \Sigma^* \times Y \rightarrow \mathbb{R}$, where Σ^* is the set possible segments and Y is the finite set of morphological features.[3] The scoring function Θ expresses the strength of association between a segment such as **ed** and a morphological feature such as Past. It is learned from a set of unsegmented word forms and their morphological label sets. We present several alternative formulations for Θ.

Using the scoring function Θ, we find the optimal segmentation $x_{max} = x_1...x_n$ of the input word form and optimal feature assignment $y_{max} = y_1 \cup ... \cup y_n$ which together maximize the value of Θ as given by Equation 1. We perform the maximization using an exact search algorithm over the set of segmentations and feature assignments. Therefore, we are guaranteed to find the optimal segmentation and feature assignment.

$$(x_{max}, y_{max}) = \underset{\substack{x_1...x_n=\mathbf{x} \\ y_1 \cup ... \cup y_n = \mathbf{y}}}{\arg\max} \sum_{x_i} \sum_{y \in y_i} \Theta(x_i, y) \quad (1)$$

We perform the maximization in Equation 1 in the following way. Let $\mathbf{x}$ be an input string and let $\mathbf{y} = \{f_1, ..., f_k\}$ be a set of morphological features. We first form an exhaustive set of segmentations of $\mathbf{x}$ (in order to allow for tractable inference, we only consider segmentations into maximally 5 segments). We then consider each segmentation $x_1...x_n = \mathbf{x}$ in turn and find the feature assignment $y_1 \cup ... \cup y_n = \mathbf{y}$ which maximizes the score Θ using a recursive algorithm presented below.

Given a partition $y_1, ..., y_n$ of a possibly empty prefix of $\mathbf{y} = \{f_1, ..., f_k\}$ (that is a collection of pointwise disjoint sets y_1, ..., y_n where $y_1 \cup ... \cup y_n = \{f_1, ..., f_j\}$ and $j \leq k$), we can find the optimal score M for assigning the remaining morphological features $\mathbf{y}_{rest} = \mathbf{y} - \{f_1, ..., f_j\}$ into sets in $y_1, ..., y_n$ using the following recursive algorithm.[4] Set $M := -\infty$ and iterate over i in $\{1, ..., n\}$.

- If $\mathbf{y}_{rest}$ is empty, then $y_1 \cup ... \cup y_n = \mathbf{y}$. If each $y_l \neq \emptyset$, assign $M := \max(M, \Theta((x_1, ..., x_n), (y_1, ..., y_n)))$.

[3] Y is finite because the inventory of morphological features is derived from a finite corpus.

[4] A natural extension of this algorithm will recover the optimal feature assignment.

- If $\mathbf{y}_{rest}$ is not empty, assign $y_i := y_i \cup \{f_{j+1}\}$ and find the optimal score M' for $\mathbf{y}'_{rest} = \{f_{j+1}, ..., f_k\} - \{f_{j+1}\}$. Set $M := \max(M, M')$.

Finally, return M.

By initially setting $y_1 = ... = y_n = \emptyset$, we can find the optimal feature assignment for the entire segmentation $x_1, ..., x_n$.

We limit the set of segmentations of a word to those which have maximally one empty substring and explore assignments where each segment is aligned with at least one morphological feature. One segment may, however, be aligned with several features. This is required when a morpheme encodes for several morphological features.

These assumptions are in line with typological considerations—agglutinative languages such as Finnish and Turkish largely associate allomorphs with a single morphological feature, while fusional languages, such as Swedish and Spanish, may associate many features with a substring (see figure 1). Allomorph overlap, where a substring **xyz** in a word has **xy** associated with one feature and **yz** with another, is generally not attested cross-linguistically which narrows down the set of hypotheses we need to consider. However, a typologically interesting case not modeled in our approach is templatic, or root-and-pattern morphology, where a discontinuous subsequence may be associated with a feature, such as in the classic Arabic example **kataba** 'to write', where root radicals associate with a stem (**ktb** = related to writing) and intervening vowels with inflectional and derivational patterns. The objective functions we develop may be adapted to this case, however, at the cost of enlarging the search space since all subsequences would need to be considered when associating parts of word forms and morphological features. Moreover, even languages with templatic morphology can be handled using the current system, provided that templatic phenomena are not annotated in the corpus. For example, a model of Arabic may represent vowel changes in the stem by declaring different stem allomorphs instead of treating the discontinuous root radical consonants as the stem, e.g. **kataba** ('to write' perfect indicative 2p masculine) vs. **taktubu** ('to write' imperfective indicative 2p masculine).

Below, we present several alternative formulations for the scoring function Θ. Two of the functions *symmetric conditional probability* and *KL-divergence* are statistics which can be computed in a straightforward manner given a training data set. The two remaining functions, the *perceptron* and *Rescorla-Wagner*, are derived by learning multi-class classifiers which predict the morphological labels of a word based on its sub-strings. The parameters of these classifiers express associations between morphological labels and substrings. We use these associations as the scoring function Θ.

3.1 Symmetric conditional probability

Intuitively, a substring x is a good candidate allomorph for a morphological feature y if x and y frequently co-occur. Symmetric conditional probability (SCP), introduced by da Silva et al. (1999) for mining of lexical multi-word units, is a mathematical realization of this principle.

The SCP of a segment x and a feature y is given by equation 2. The probability $p(x)$ is the frequency of words having substring x, $p(y)$ the frequency of words having morphological label y and $p(x, y)$ the frequency of words having both substring x and label y.

$$\mathrm{SCP}(x, y) = p(x|y)p(y|x) = \frac{p(x, y)^2}{p(x)p(y)} \quad (2)$$

By setting $\Theta(x, y) = \mathrm{SCP}(x, y)$, we can use the symmetric conditional probability as a scoring function.

3.2 Perceptron

We explore a simple extension of the classical perceptron learning algorithm (Rosenblatt, 1958) for multi-label classification (Tsoumakas and Katakis, 2007). Instead of predicting a single label for each input instance, we predict a set of outputs corresponding to the morphological features related to a word.

We start with a standard perceptron classifier defined by a feature extraction function $f : \Sigma^* \rightarrow \{0, 1\}^k$, which maps word forms x into a vector, and one parameter vector $\phi_y \in \mathbb{R}^k$ for each morphological feature $y \in Y$. Here, k is the total number of distinct substrings that occur in the data set $\mathcal{D}$.[5] Intuitively, f extracts substrings of x. More formally, it maps a word form x into a vector in a space where each dimension corresponds to a string in Σ^* and $f(x)[i] = 1$, iff x has a substring corresponding to the ith dimension. Inference in

[5] k is between 144,000 and 423,000 for all of the data sets considered in this paper.

the model is defined by Equation 3 and parameter updates are defined by Equation 4, where y_{gold} is the gold standard label.

$$y_{max} = \arg\max_{y \in Y} \phi_y^\top f(x) \tag{3}$$

$$\phi_{y_{max}} := \phi_{y_{max}} - f(x) \text{ and } \phi_{y_{gold}} := \phi_{y_{gold}} + f(x) \tag{4}$$

We modify standard perceptron updates in the following way: For a word x with a set of morphological labels Y, where $|Y| = n$, we examine the set N consisting of the top-n labels returned by the perceptron classifier using the current parameter estimates. We then perform a negative update for parameters corresponding to labels which were not associated with the word form x in the gold standard, that is labels in the set $N - Y$. Conversely, we perform a positive update for parameters which were associated with word form x, that is morphological features in the set $Y - N$.

Clearly, we perform no updates for a particular word form x, iff the top-n candidates returned by the classifier exactly correspond to the set of morphological features of x.

We first train a system using the aforementioned variant of the perceptron algorithm. As feature templates, we us the substrings of words in $\mathcal{D}$. We then use the parameter values corresponding to associations of substrings x and features y, learned by the perceptron algorithm, as scores $\Theta(x, y)$.

3.3 Rescorla-Wagner learning

The Rescorla-Wagner (R-W) rule (Rescorla and Wagner, 1972) is a model of classical conditioning that provides an account of the association strength between a conditioned stimulus (CS) and an unconditioned stimulus (US); or, alternatively, the strength between a *stimulus* and the expectation of *reward*. This type of a learning model has been applied to acquisition of plurals (Ramscar and Yarlett, 2007; Ramscar, 2013), number names (Ramscar et al., 2011) word recognition (Baayen et al., 2011), and typology of number encoding in inflectional morphology (Ackerman et al., 2016).

In the single stimulus case, we have an expected reward v which is calculated as a linear combination of a binary stimulus u and a learned weight w:

$$v = wu \tag{5}$$

As stimuli arrive possibly paired with a reward, the weight w is updated depending on whether the reward was present as $w = w + \varepsilon\delta u$, where δ is set in proportion to r, the 'actual' reward, usually set to 1 or 100 if the association is valid, else 0. The quantity $\delta = r - v$ is hence the prediction error which drives association updates toward 0 or toward the maximum association score and ε is a learning rate (set to 0.01 here). In our model, the *reward* is a morphological label, and the *stimuli* are the substrings present in the word forms witnessed. We extend the common single-stimulus/single-reward R-W model to one which learns association strengths of multiple stimuli and multiple rewards in a standard way (Dayan and Abbott, 2001). Each possible morphological label is associated with a weight vector $\mathbf{w}$ where each dimension corresponds to a string in Σ^*, as in the perceptron case. As stimuli arrive, weight updates are performed per label as:

$$\mathbf{w} = \mathbf{w} + \varepsilon\delta\mathbf{u} \text{ where } \delta = r - v \tag{6}$$

Here, as before, r is 0 if the label is absent and 100 if it is present.

Learning is very similar to perceptron learning—the conditions under which R-W learning and perceptron training is identical is explored in detail in Dawson (2008). The main difference between our R-W and perceptron implementations is that perceptron updates are only performed if the n labels present in a word form do not appear in the n-best scoring list, while R-W updates are always done if the expectation produced by summing the individual expectations caused by the substrings in a word fails to match the maximum label 'reward'.

3.4 Kullback-Leibler divergence

Given a set of labeled words $U \subset \mathcal{D}$ in our data set, we can examine the distribution of morphological features in U defined by $p(f|U) \propto |\{(x, y) \in U | f \in y\}|$ for all $f \in Y$. We can examine different subsets of $\mathcal{D}$ defined by criteria concerning (1) morphological features, or (2) existence of a given substring in word forms. Intuitively, a substring s is a good candidate morpheme for a morphological feature f, if $U_s = \{(x, y) \in \mathcal{D} | s \text{ is a substring of } x\}$ and $U_f = \{(x, y) \in \mathcal{D} | f \in y\}$ define similar distributions of morphological features.

Kullback-Leibler (KL) divergence is a widely

	# train wf	# test wf	# lemmas	feat. types
Eng	10,000	300	8591	7
Fin	12,693	300	10049	39
Swe	10,000	300	6589	23
Tur	7,645	300	2523	36

Table 1: Data set sizes for English, Finnish, Swedish and Turkish.

used measure for the distance of two discrete probability distributions defined on the same sample space defined by Equation 7.

$$\mathrm{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{7}$$

We use the negative KL divergence of the distributions over morphological features defined by a substring x and a morphological feature f, respectively, as score $\Theta(x, f)$.

3.5 Baseline

As our baseline, we use the unsupervised morphological segmentation given by the Morfessor Baseline method (Creutz and Lagus, 2005). We use its default settings for all parameters.

We assign labels to segments based on maximum likelihood as defined by co-occurrence of segments and labels in the segmented data set. When there are fewer segments than morphological features, we assign at least one feature per segment. Otherwise, we assign at most one feature per segment while maximizing the joint probability of segments and morphological features.

This baseline was chosen because it is easily accessible to most researchers and very easy and fast to apply.

4 Data

We use data from the 2016 SIGMORPHON shared task on morphological re-inflection (Cotterell et al., 2016b) (Finnish and Turkish) and the 2017 CoNLL shared task on morphological re-inflection[6] (English and Swedish). Figure 2 shows an example of the data format and Table 1 shows details for each data set.

We use the training and test sets from subtask 1 from the SIGMORPHON shared task (Cotterell et al., 2016a) and the task 1 high training set together with the task 1 development set from the CoNLL shared task 2017 (Cotterell et al., 2017). Figure 3 shows an example of the annotated test data.

[6] https://sites.google.com/view/conll-sigmorphon2017/

5 Experiments

We first train each of the scoring functions presented in Section 3 on the combined training data and *unsegmented* test data. After that, we find the optimal segmentation and label alignment for each word in the test data using each scoring function. We explore all segmentations consisting of maximally 5 segments and all assignments of morphological features to the segments using a dynamic algorithm in order to speed up inference.

For perceptron and R-W learning, we run the training algorithm for three epochs over the train and test data. The learning rate ε for R-W learning is fixed to 0.01 and the maximum possible association response r is fixed to 100.

We evaluate each scoring function with regard to three different criteria: (1) identification of morpheme boundaries, (2) identification of unlabeled morphemes, and (3) identification of labeled morphemes. For each criterion, we give recall, precision and F_1-score. Evaluation criteria (1) and (2) are very similar, but we include both for easier comparison with earlier work in the field of unsupervised morphological segmentation.

To illustrate our evaluation scheme, consider the following gold standard segmentation and alignment for English

```
ping/ping ing/V.PTCP,PRS NULL/V
```

The aligned form contains three morpheme boundaries: at index 1 (start of word), at index 4 (between the stem and participle suffix) and at index 7 (end of word). It contains two unlabeled morphemes: **ping** and **ing**, and four labeled morphemes: **ping/ping**, **ing/V.PTCP**, **ing/PRS** and **NULL/V**. Counts for these units are used to compute recall, precision and F_1-score for each evaluation criterion.

6 Results

Table 2 shows the results of all experiments for each language.

In general, each of the scoring functions performs substantially better than the baseline Morfessor system. However, SCP delivers lower unlabeled morpheme F-scores for Turkish and KL-divergence gives lower performance on morpheme boundary detection for Finnish.

psychoanalyse	V,V.PTCP,PRS	psychoanalysing
aalloittaisuus	pos=N,case=ON+ESS,num=PLN	aalloittaisuuksilla
centralbank	N,DEF,GEN,SG	centralbankens
haberleşmek	V,IND,3,SG,PST,PROG,POS,DECL	haberleşiyordu

Figure 2: Example lines from the English, Finnish, Swedish and Turkish training data set of. The first field contains the lemma, the second field contains additional morphological features and the last field contains the word form.

```
autofocuss/autofocus ed/PST NULL/V
paali/paali n/case=ACC NULL/num=SGN,pos=N
kammarrätt/kammarrätt s/GEN NULL/N,INDF,SG
âmâ/âmâ lar/PL dan/ABL NULL/N
```

Figure 3: Example entries from the annotated English, Finnish, Swedish and Turkish test sets. We align the stem with the lemma and the part-of-speech with the zero allomorph.

R-W seems to deliver consistently competitive performance when compared to the other systems on morpheme boundary recovery and morpheme identification. The performance of the perceptron algorithm is quite similar to the R-W but, in general, lower. The perceptron algorithm, however, delivers the best performance for Turkish.

KL divergence seems to perform the worst of all of the scoring functions. It delivers markedly worse performance on the Swedish data set than the other systems.

SCP delivers superior performance when compared to R-W for Finnish on morpheme boundary recovery and morpheme identification. However, its performance on English and and Turkish is substantially worse than both R-W and the perceptron algorithm.

In the case of labeled morphemes, SCP seems to deliver consistently good performance. It outperforms R-W even in the case of Turkish and English, where it delivers substantially worse performance on unlabeled morpheme identification.

7 Discussion

The results show clear improvement over the baseline approach of first applying unsupervised morphological segmentation and then assigning labels based on co-occurrence counts of segments and labels. That is, including information about morphological features in the segmentation process is clearly beneficial.

The perceptron and R-W learning algorithms have very similar performance, which can be explained by the fact that the algorithms themselves are quite similar. However, the R-W algorithm seems to deliver somewhat superior performance. One possible reason for this is that the R-W will prefer solutions where one substring in the word explains one morphological feature, whereas the perceptron algorithm does not have such a bias. This can be attributed to the 'blocking' effect of R-W learning: when one feature (substring) has already been weighted early during training enough to yield a maximum response (label), no updates are made for other features which may also co-occur with the same label.

The fact that both R-W and the perceptron algorithm seem to perform poorly for labeled morpheme identification can be explained by the fact that both algorithms are trained to predict each of the morphological labels of the word from all substrings occurring in the word. This can lead to confusion of features for morphemes occurring in the same word. For example, the R-W performs comparatively poorly on labeled morpheme identification for the Finnish, Swedish and English data sets. This happens because it assigns the part-of-speech feature to the stem in many words but the gold standard analysis is that the part-of-speech is aligned with the zero morpheme. Conversely, it also assigns the lemma to the zero morpheme in many words, whereas the gold standard instead assigns lemmas to stems. Note that the decision to align part-of-speech with the zero morpheme instead of the word stem is fairly arbitrary. Therefore, a different gold standard segmentation could give substantially higher labeled morpheme performance for R-W.

The arbitrariness of the gold standard annotation as regards certain features may be avoided by a different evaluation scheme where no gold standard is used. One can, for example, leave a held-out data set and first segment and label the data on a training section, and then investigate

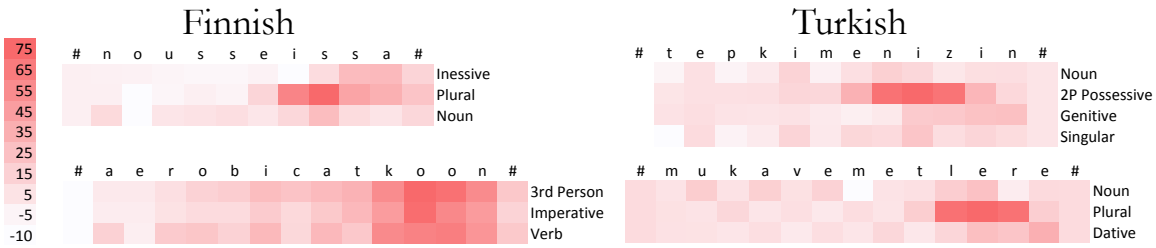

Figure 4: Example activations for two inflected Finnish words (**noussut** 'risen', **aerobicata** 'to do aerobics') and two Turkish words (**tepkime** 'reaction', **mukavemet** 'durability') with Rescorla-Wagner learning. The activation score at each character is calculated as a sum of the activations associated with the substrings that the character participates in. Standard linguistic analyses have the Finnish inessive as **-ssa**, the plural as **-i-**, and both the imperative and 3rd person fused as **-koon**. For Turkish, the 2P Possessive is **-niz-**, the genitive is **-in**, the plural is **-ler-**, and the dative is **-e**.

(a)

	Eng	**Fin**	**Swe**	**Tur**
Kullback-Leibler Divergence				
R	93.91	82.74	73.81	81.25
P	87.15	80.36	65.89	76.60
F_1	90.41	81.54	69.63	78.86
Perceptron				
R	98.81	80.67	86.15	86.68
P	95.06	88.05	76.54	90.54
F_1	96.90	84.20	81.06	**88.57**
Rescorla-Wagner				
R	98.93	83.74	82.58	82.88
P	97.87	86.81	82.22	91.96
F_1	**98.40**	85.23	82.40	87.18
Symmetric Conditional Probability				
R	89.02	81.56	77.38	70.65
P	95.15	91.58	94.33	90.28
F_1	91.99	**86.28**	**85.02**	79.27
Morfessor baseline				
R	80.79	67.36	76.19	77.26
P	61.10	65.67	72.35	93.22
F_1	69.58	66.50	74.22	84.50

(b)

	Eng	**Fin**	**Swe**	**Tur**
Kullback-Leibler Divergence				
R	69.52	45.66	15.71	44.28
P	62.02	43.82	13.35	40.97
F_1	65.56	44.72	14.43	42.56
Perceptron				
R	90.15	47.62	44.55	61.32
P	84.94	54.05	37.57	65.13
F_1	87.47	50.63	40.76	**63.16**
Rescorla-Wagner				
R	94.98	50.84	43.91	56.84
P	93.42	53.53	43.63	65.76
F_1	**94.19**	52.16	**43.77**	60.97
Symmetric Conditional Probability				
R	64.31	50.14	37.82	27.99
P	71.49	59.37	51.53	39.89
F_1	67.71	**54.37**	43.62	32.89
Morfessor baseline				
R	21.19	9.94	21.79	39.93
P	14.14	9.59	20.27	52.20
F_1	16.96	9.77	21.00	45.24

(c)

	Eng	**Fin**	**Swe**	**Tur**
Kullback-Leibler Divergence				
R	74.18	39.70	8.88	31.36
P	74.11	37.07	7.64	27.09
F_1	74.15	38.34	8.22	29.07
Perceptron				
R	90.06	34.77	30.30	59.91
P	90.06	33.59	27.30	54.70
F_1	**90.06**	34.17	28.72	**57.19**
Rescorla-Wagner				
R	43.96	29.98	24.31	43.14
P	43.96	29.98	24.96	43.58
F_1	43.96	29.98	24.63	43.36
Symmetric Conditional Probability				
R	66.37	56.45	62.88	52.81
P	66.37	53.90	64.07	53.35
F_1	66.37	**55.15**	**63.47**	53.08
Morfessor baseline				
R	1.77	9.66	20.11	8.74
P	1.76	10.97	25.09	11.21
F_1	1.77	10.27	22.32	9.82

Table 2: Results for (a) morpheme boundaries; (b) unlabeled morphemes; (c) labeled morphemes. For each language and each task, the scoring function delivering the best performance is shown in boldface.

how many new allomorphs are implicitly detected when the held-out data is also segmented and labeled. The expectation is that very few new allomorphs should be found in a held-out set if a model assigns labels to substrings in a consistent manner. Whether a good score on such an evaluation would correspond to linguistically motivated allomorph sets is a question we intent to investigate in the future. If so, a robust evaluation could potentially be made without any gold segmentation and labeling at all.

In addition to the scoring functions presented in Section 3, we investigated a number of other scoring functions, for example pointwise mutual information (PMI)[7] of segments and morphological features, but these did not yield competitive performance according to preliminary experiments. We also experimented with IBM models (Brown et al., 1993) for alignment of characters to morphosyntactic labels, which also performed poorly.

SCP performs poorly on the English and Turkish data sets. For English, a major problem is that SCP does not find the present participle suffix **ing**. This suffix is problematic because it is associated with a combination of morphological features, namely present tense and the participle feature. Both of these co-occur more frequently with other suffixes (**ed** in the case of participle and **NULL** in the case of present tense), however, when they co-occur, they always occur with the **ing** suffix. This seems to be a problem for SCP which encodes a strong preference that there be a one-to-one mapping between morphemes and features.

A possible explanation for the poor performance of SCP on the Turkish data set is that this is the smallest of all data sets, while still having a very large number of morphological features.

In this investigation, we have not exhausted the set of reasonable scoring functions. One objective function that is particularly interesting is to simply try to minimize the total number of different allomorphs discovered in the data. This objective function is difficult to integrate in our current approach since the function is discontinuous. In essence, this objective function calls for an algorithm that discovers a segmentation and labeling of the data such that the sum total of different allomorph types is minimized. The problem appears to be computationally intractable in principle, since it bears strong similarities to other intractable problems such as set covering. But good heuristic solvers for NP-complete problems such as Moskewicz et al. (2001) may perhaps be harnessed to find good solutions under this formulation. A thorough analysis and evaluation of this type of model remains future work.

[7] The reason for the poor performance of PMI is that it will often align features with rare substrings and, therefore, it can assign a great number of distinct allomorphs to the same morphological feature. To illustrate this, let $\mathrm{pmi}(x, y) = \log p(x, y)/(p(x)p(y))$ be the PMI of segment x and feature y. This quantity can never exceed $\log 1/p(y)$ because $p(x, y) \leq p(x)$. Assume that x only occurs once in the training corpus and the sole occurrence is in a word with feature y. Thus $p(x, y) = p(x)$ and $\mathrm{pmi}(x, y) = \log 1/p(y)$, i.e. the maximal PMI for any segment x given feature y.

8 Conclusion

We have presented a new learning problem for natural language processing, namely weakly supervised learning of allomorphy. The problem is important from a practical point of view because there are many morphologically annotated corpora where the annotation is not extended to the morpheme level. It is also relevant from a theoretical point of view because it is related to L1 morphology learning.

We explored four different learning methods: KL divergence, perceptron learning, R-W learning and SCP. We compared these to a baseline consisting of unsupervised morphological segmentation augmented by a straightforward labeling mechanism. Our results show that weak supervision delivers sizable improvements when evaluated with regard to F_1-score on labeled and unlabeled segmentation. According to our experiments, R-W learning, while not only efficient, also delivers the best results on this task.

Acknowledgements

We wish to thank Jordan Boyd-Graber, Koen Claessen, Gerlof Bouma, and Hubie Chen for very helpful discussion and comments. This work has been supported in part by the Defense Advanced Research Projects Agency (DARPA) in the program Low Resource Languages for Emergent Incidents (LORELEI).

References

Farrell Ackerman, Robert Malouf, and James P. Blevins. 2016. Patterns and discriminability in language analysis. *Word Structure* 9(2):132–155.

R. Harald Baayen, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review* 118(3):438–481.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–313.

Eve V. Clark and Ruth A. Berman. 1984. Structure and use in the acquisition of word formation. *Language* pages 542–590.

Eve V. Clark and Barbara Frant Hecht. 1982. Learning to coin agent and instrument nouns. *Cognition* 12(1):1–24.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Association for Computational Linguistics, Vancouver, Canada.

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*. Association for Computational Linguistics, Berlin, Germany, pages 10–22.

Ryan Cotterell, Thomas Müller, Alexander M. Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-Markov models. In *CoNLL*. pages 164–174.

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*. Association for Computational Linguistics, San Diego, California, pages 664–669.

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*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 113–132.

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*. Association for Computational Linguistics, Rochester, New York, pages 155–163.

Michael R. W. Dawson. 2008. Connectionism and classical conditioning. *Comparative Cognition and Behavior Reviews* 3:1–115.

P. Dayan and L. F. Abbott. 2001. Classical conditioning and reinforcement learning. *Theoretical Neuroscience* .

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

Manaal Faruqui, Ryan McDonald, and Radu Soricut. 2016. Morpho-syntactic lexicon generation using graph-based semi-supervised learning. *Transactions of the Association for Computational Linguistics* 4:1–16.

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

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*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 1177–1185.

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*. Association for Computational Linguistics, Austin, Texas, pages 961–967.

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)*. Association for Computational Linguistics, pages 78–86.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2):313–330.

Matthew W. Moskewicz, Conor F. Madigan, Ying Zhao, Lintao Zhang, and Sharad Malik. 2001. Chaff: Engineering an efficient SAT solver. In *Proceedings of the 38th annual Design Automation Conference*. ACM, pages 530–535.

Elissa L. Newport. 2016. Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition* 8(03):447–461.

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*. Association for Computational Linguistics, pages 209–217.

Michael Ramscar. 2013. Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija* 46(4):377–396.

Michael Ramscar, Melody Dye, Hanna Muenke Popick, and Fiona O'Donnell-McCarthy. 2011. The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PloS one* 6(7).

Michael Ramscar and Daniel Yarlett. 2007. Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science* 31(6):927–960.

Patricia A. Reeder, Elissa L. Newport, and Richard N. Aslin. 2013. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology* 66(1):30–54.

Robert A. Rescorla and Allan R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory* 2:64–99.

Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6):386–408.

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.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294):1926–1928.

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*. Association for Computational Linguistics, pages 67–72.

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*. Association for Computational Linguistics, Denver, Colorado, pages 1627–1637.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *IJDWM* 3(3):1–13.