Character-Level Models versus Morphology in Semantic Role Labeling

Abstract

Character-level models have become a popular approach specially for their accessibility and ability to handle unseen data. However, little is known on their ability to reveal the underlying morphological structure of a word, which is a crucial skill for high-level semantic analysis tasks, such as semantic role labeling (SRL). In this work, we train various types of SRL models that use word, character and morphology level information and analyze how performance of characters compare to words and morphology for several languages. We conduct an in-depth error analysis for each morphological typology and analyze the strengths and limitations of character-level models that relate to out-of-domain data, training data size, long range dependencies and model complexity. Our exhaustive analyses shed light on important characteristics of character-level models and their semantic capability.

Introduction

Encoding of words is perhaps the most important step towards a successful end-to-end natural language processing application. Although word embeddings have been shown to provide benefit to such models, they commonly treat words as the smallest meaning bearing unit and assume that each word type has its own vector representation. This assumption has two major shortcomings especially for languages with rich morphology: (1) inability to handle unseen or out-of-vocabulary (OOV) word-forms (2) inability to exploit the regularities among word parts. The limitations of word embeddings are particularly pronounced in sentence-level semantic tasks, especially in languages where word parts play a crucial role. Consider the Turkish sentences “Köy+lü-ler (villagers) şehr+e (to town) geldi (came)” and “Sendika+lı-lar (union members) meclis+e (to council) geldi (came)”. Here the stems köy (village) and sendika (union) function similarly in semantic terms with respect to the verb come (as the origin of the agents of the verb), where

şehir (town) and meclis (council) both function as the end point. These semantic similarities are determined by the common word parts shown in bold. However ortographic similarity does not always correspond to semantic similarity. For instance the ortographically similar words knight and night have large semantic differences. Therefore, for a successful semantic application, the model should be able to capture both the regularities, i.e, morphological tags and the irregularities, i.e, lemmas of the word.

Morphological analysis already provides the aforementioned information about the words. However access to useful morphological features may be problematic due to software licensing issues, lack of robust morphological analyzers and high ambiguity among analyses. Character-level models (CLM), being a cheaper and accessible alternative to morphology, have been reported as performing competitively on various NLP tasks BIBREF0 , BIBREF1 , BIBREF2 . However the extent to which these tasks depend on morphology is small; and their relation to semantics is weak. Hence, little is known on their true ability to reveal the underlying morphological structure of a word and their semantic capabilities. Furthermore, their behaviour across languages from different families; and their limitations and strengths such as handling of long-range dependencies, reaction to model complexity or performance on out-of-domain data are unknown. Analyzing such issues is a key to fully understanding the character-level models.

To achieve this, we perform a case study on semantic role labeling (SRL), a sentence-level semantic analysis task that aims to identify predicate-argument structures and assign meaningful labels to them as follows:

$[$ Villagers $]$ comers came $[$ to town $]$ end point

We use a simple method based on bidirectional LSTMs to train three types of base semantic role labelers that employ (1) words (2) characters and character sequences and (3) gold morphological analysis. The

gold morphology serves as the upper bound for us to compare and analyze the performances of character-level models on languages of varying morphological typologies. We carry out an exhaustive error analysis for each language type and analyze the strengths and limitations of character-level models compared to morphology. In regard to the diversity hypothesis which states that diversity of systems in ensembles lead to further improvement, we combine character and morphology-level models and measure the performance of the ensemble to better understand how similar they are.

We experiment with several languages with varying degrees of morphological richness and typology: Turkish, Finnish, Czech, German, Spanish, Catalan and English. Our experiments and analysis reveal insights such as:

Method

Formally, we generate a label sequence $\vec{l}$ for each sentence and predicate pair: (s,p) . Each $l_t\in \vec{l}$ is chosen from $\mathcal {L}=\lbrace \mathit {roles \cup nonrole}\rbrace $, where $roles$ are language-specific semantic roles (mostly consistent with PropBank) and $nonrole$ is a symbol to present tokens that are not arguments. Given $\theta $ as model parameters and g_t as gold label for t_{th} token, we find the parameters that minimize the negative log likelihood of the sequence:

$$\hat{\theta }=\underset{\theta }{\arg \min } \left(-\sum _{t=1}^n log (p(g_t|\theta ,s,p)) \right)$$ (Eq. 7)

Label probabilities, $p(l_t|\theta ,s,p)$, are calculated with equations given below. First, the word encoding layer splits tokens into subwords via $\rho $ function.

$$\rho (w) = {s_0,s_1,..,s_n}$$ (Eq. 8)

As proposed by BIBREF0 , we treat words as a sequence of subword units. Then, the sequence is fed to a simple bi-LSTM network BIBREF15 , BIBREF16 and hidden states from each direction are weighted with a set of parameters which are also learned during training. Finally, the weighted vector is used as the word embedding given in Eq. 9 .

$$hs_f, hs_b = \text{bi-LSTM}({s_0,s_1,..,s_n}) \\
\vec{w} = W_f \cdot hs_f + W_b \cdot hs_b + b$$ (Eq. 9)

There may be more than one predicate in the sentence so it is crucial to inform the network of which arguments we aim to label. In order to mark the predicate of interest, we concatenate a predicate flag pf_t to the word embedding vector.

$$\vec{x_{t}} = [\vec{w};pf_t]$$ (Eq. 10)

Final vector, $\vec{x_t}$ serves as an input to another bi-LSTM unit.

$$\vec{h_{f}, h_{b}} = \text{bi-LSTM}(x_{t})$$ (Eq. 11)

Finally, the label distribution is calculated via softmax function over the concatenated hidden states from both directions.

$$\vec{p(l_t|s,p)} = softmax(W_{l}\cdot [\vec{h_{f}};\vec{h_{b}}]+\vec{b_{l}})$$ (Eq. 12)

For simplicity, we assign the label with the highest probability to the input token. .

Subword Units

We use three types of units: (1) words (2) characters and character sequences and (3) outputs of morphological analysis. Words serve as a lower bound; while morphology is used as an upper bound for comparison. Table 1 shows sample outputs of various $\rho $ functions.

Here, char function simply splits the token into its characters. Similar to n-gram language models, char3 slides a character window of width $n=3$ over the token. Finally, gold morphological features are used as outputs of morph-language. Throughout this paper, we use morph and oracle interchangably, i.e., morphology-level models (MLM) have access to gold tags unless otherwise is stated. For all languages, morph outputs the lemma of the token followed by language specific morphological tags. As an exception, it outputs additional information for some languages, such as parts-of-speech tags for Turkish. Word segmenters such as Morfessor and Byte Pair Encoding (BPE) are other commonly used subword units. Due to low scores obtained from our preliminary experiments and unsatisfactory results from previous studies BIBREF13 , we excluded these units.

Experiments

We use the datasets distributed by LDC for Catalan (CAT), Spanish (SPA), German (DEU), Czech (CZE) and English (ENG) BIBREF17 , BIBREF18 ; and datasets made available by BIBREF19 , BIBREF20 for Finnish (FIN) and Turkish (TUR) respectively . Datasets are provided with syntactic dependency annotations and semantic roles of verbal predicates. In addition, English supplies nominal predicates annotated with semantic roles and does not provide any morphological feature.

Statistics for the training split for all languages are given in Table 2 . Here, #pred is number of predicates, and #role refers to number distinct semantic roles that occur more than 10 times. More detailed statistics about the datasets can be found in BIBREF27 , BIBREF19 , BIBREF20 .

Experimental Setup

To fit the requirements of the SRL task and of our model, we performed the following:

Multiword expressions (MWE) are represented as a single token, (e.g., Confederación_Francesa_del_Trabajo), that causes notably long character sequences which are hard to handle by LSTMs. For the sake of memory efficiency and performance, we used an abbreviation (e.g., CFdT) for each MWE during training and testing.

Original dataset defines its own format of semantic annotation, such as 17:PBArgM_mod $\mid $ 19:PBArgM_mod meaning the node is an argument of 17_{th} and 19_{th} tokens with ArgM-mod (temporary modifier) semantic role. They have been converted into CoNLL-09 tabular format, where each predicate's arguments are given in a specific column.

Words are splitted from derivational boundaries in the original dataset, where each inflectional group is represented as a separate token. We first merge boundaries of the same word, i.e, tokens of the word, then we use our own $\rho $ function to split words into subwords.

We lowercase all tokens beforehand and place special start and end of the token characters. For all experiments, we initialized weight parameters orthogonally and used one layer bi-LSTMs both for subword composition and argument labeling with hidden size of 200. Subword embedding size is chosen as 200. We used gradient clipping and early stopping to prevent overfitting. Stochastic gradient descent is used as the optimizer. The initial learning rate is set to 1 and reduced by half if scores on development set do not improve after 3 epochs. We use the provided splits and evaluate the results with the official evaluation script provided by CoNLL-09 shared task. In this work (and in most of the recent SRL works), only the scores for argument labeling are reported, which may cause confusions for the readers while

comparing with older SRL studies. Most of the early SRL work report combined scores (argument labeling with predicate sense disambiguation (PSD)). However, PSD is considered a simpler task with higher F1 scores . Therefore, we believe omitting PSD helps us gain more useful insights on character level models.

Results and Analysis

Our main results on test and development sets for models that use words, characters (char), character trigrams (char3) and morphological analyses (morph) are given in Table 3 . We calculate improvement over word (IOW) for each subword model and improvement over the best character model (IOC) for the morph. IOW and IOC values are calculated on the test set.

The biggest improvement over the word baseline is achieved by the models that have access to morphology for all languages (except for English) as expected. Character trigrams consistently outperformed characters by a small margin. Same pattern is observed on the results of the development set. IOW has the values between 0% to 38% while IOC values range between 2%-10% dependending on the properties of the language and the dataset. We analyze the results separately for agglutinative and fusional languages and reveal the links between certain linguistic phenomena and the IOC, IOW values.

Similarity between models

One way to infer similarity is to measure diversity. Consider a set of baseline models that are not diverse, i.e., making similar errors with similar inputs. In such a case, combination of these models would not be able to overcome the biases of the learners, hence the combination would not achieve a better result. In order to test if character and morphological models are similar, we combine them and measure the performance of the ensemble. Suppose that a prediction p_{i} is generated for each token by a model

m_i , $i \in n$, then the final prediction is calculated from these predictions by:

$$p_{final} = f(p_0, p_1,..,p_n|\phi)$$ (Eq. 36)

where f is the combining function with parameter $\phi $. The simplest global approach is averaging (AVG), where f is simply the mean function and p_i s are the log probabilities. Mean function combines model outputs linearly, therefore ignores the nonlinear relation between base models/units. In order to exploit nonlinear connections, we learn the parameters $\phi $ of f via a simple linear layer followed by sigmoid activation. In other words, we train a new model that learns how to best combine the predictions from subword models. This ensemble technique is generally referred to as stacking or stacked generalization (SG).

Although not guaranteed, diverse models can be achieved by altering the input representation, the learning algorithm, training data or the hyperparameters. To ensure that the only factor contributing to the diversity of the learners is the input representation, all parameters, training data and model settings are left unchanged.

Our results are given in Table 4 . IOB shows the improvement over the best of the baseline models in the ensemble. Averaging and stacking methods gave similar results, meaning that there is no immediate nonlinear relations between units. We observe two language clusters: (1) Czech and agglutinative languages (2) Spanish, Catalan, German and English. The common property of that separate clusters are (1) high OOV% and (2) relatively low OOV%. Amongst the first set, we observe that the improvement gained by character-morphology ensembles is higher (shown with green) than ensembles between characters and character trigrams (shown with red), whereas the opposite is true for the second set of languages. It can be interpreted as character level models being more similar to the morphology level models for the first cluster, i.e., languages with high OOV%, and characters and morphology being more

diverse for the second cluster.

Limitations and Strengths

To expand our understanding and reveal the limitations and strengths of the models, we analyze their ability to handle long range dependencies, their relation with training data and model size; and measure their performances on out of domain data.

Long Range Dependencies

Long range dependency is considered as an important linguistic issue that is hard to solve. Therefore the ability to handle it is a strong performance indicator. To gain insights on this issue, we measure how models perform as the distance between the predicate and the argument increases. The unit of measure is number of tokens between the two; and argument is defined as the head of the argument phrase in accordance with dependency-based SRL task. For that purpose, we created bins of [0-4], [5-9], [10-14] and [15-19] distances. Then, we have calculate F1 scores for arguments in each bin. Due to low number of predicate-argument pairs in buckets, we could not analyze German and Turkish; and also the bin [15-19] is only used for Czech. Our results are shown in Fig. 3 . We observe that either char or char3 closely follows the oracle for all languages. The gap between the two does not increase with the distance, suggesting that the performance gap is not related to long range dependencies. In other words, both characters and the oracle handle long range dependencies equally well.

Training Data Size

We analyzed how char3 and oracle models perform with respect to the training data size. For that purpose, we trained them on chunks of increasing size and evaluate on the provided test split. We used

units of 2000 sentences for German and Czech; and 400 for Turkish. Results are shown in Fig. 4 . Apparently as the data size increases, the performances of both models logarithmically increase - with a varying speed. To speak in statistical terms, we fit a logarithmic curve to the observed F1 scores (shown with transparent lines) and check the x coefficients, where x refers to the number of sentences. This coefficient can be considered as an approximation to the speed of growth with data size. We observe that the coefficient is higher for char3 than oracle for all languages. It can be interpreted as: in the presence of more training data, char3 may surpass the oracle; i.e., char3 relies on data more than the oracle.

Out-of-Domain (OOD) Data

As part of the CoNLL09 shared task BIBREF27 , out of domain test sets are provided for three languages: Czech, German and English. We test our models trained on regular training dataset on these OOD data. The results are given in Table 5 . Here, we clearly see that the best model has shifted from oracle to character based models. The dramatic drop in German oracle model is due to the high lemma OOV rate which is a consequence of keeping compounds as a single lemma. Czech oracle model performs reasonably however is unable to beat the generalization power of the char3 model. Furthermore, the scores of the character models in Table 5 are higher than the best OOD scores reported in the shared task BIBREF27 ; even though our main results on evaluation set are not (except for Czech). This shows that character-level models have increased robustness to out-of-domain data due to their ability to learn regularities among data.

Model Size

Throughout this paper, our aim was to gain insights on how models perform on different languages rather than scoring the highest F1. For this reason, we used a model that can be considered small when compared to recent neural SRL models and avoided parameter search. However, we wonder how the

models behave when given a larger network. To answer this question, we trained char3 and oracle models with more layers for two fusional languages (Spanish, Catalan), and two agglutinative languages (Finnish, Turkish). The results given in Table 6 clearly shows that model complexity provides relatively more benefit to morphological models. This indicates that morphological signals help to extract more complex linguistic features that have semantic clues.

Predicted Morphological Tags

Although models with access to gold morphological tags achieve better F1 scores than character models, they can be less useful a in real-life scenario since they require gold tags at test time. To predict the performance of morphology-level models in such a scenario, we train the same models with the same parameters with predicted morphological features. Predicted tags were only available for German, Spanish, Catalan and Czech. Our results given in Fig. 5 , show that (except for Czech), predicted morphological tags are not as useful as characters alone.

Conclusion

Character-level neural models are becoming the defacto standard for NLP problems due to their accessibility and ability to handle unseen data. In this work, we investigated how they compare to models with access to gold morphological analysis, on a sentence-level semantic task. We evaluated their quality on semantic role labeling in a number of agglutinative and fusional languages. Our results lead to the following conclusions:

Acknowledgements

Gözde Gül Şahin was a PhD student at Istanbul Technical University and a visiting research student at

University of Edinburgh during this study. She was funded by Tübitak (The Scientific and Technological Research Council of Turkey) 2214-A scholarship during her visit to University of Edinburgh. She was granted access to CoNLL-09 Semantic Role Labeling Shared Task data by Linguistic Data Consortium (LDC). This work was supported by ERC H2020 Advanced Fellowship GA 742137 SEMANTAX and a Google Faculty award to Mark Steedman. We would like to thank Adam Lopez for fruitful discussions, guidance and support during the first author's visit.