Character-Level Models versus Morphology in Semantic Role Labeling

Gözde Gül Şahin
Department of Computer Science
Technische Universität Darmstadt
Darmstadt, Germany
`isguderg@itu.edu.tr`

Mark Steedman
School of Informatics
University of Edinburgh
Edinburgh, Scotland
`steedman@inf.ed.ac.uk`

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.

1 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 (Ling et al., 2015; Plank et al., 2016; Lee et al., 2017). 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:

- CLMs provide great improvements over whole-word-level models despite not being able to match the performance of morphology-level models (MLMs) for *in-domain* datasets. However their performance surpass all MLMs on *out-of-domain* data,

- Limitations and strengths differ by morphological typology. Their limitations for agglutinative languages are related to rich *derivational morphology* and high *contextual ambiguity*; whereas for fusional languages they are related to *number of morphological tags* (morpheme ambiguity) ,

- CLMs can handle long-range dependencies equally well as MLMs,

- In presence of more training data, CLM's performance is expected to improve faster than of MLM.

2 Related Work

Neural SRL Methods: Neural networks have been first introduced to the SRL scene by Collobert et al. (2011), where they use a unified end-to-end convolutional network to perform various NLP tasks. Later, the combination of neural networks (LSTMs in particular) with traditional SRL features (categorical and binary) has been introduced (FitzGerald et al., 2015). Recently, it has been shown that careful design and tuning of deep models can achieve state-of-the-art with no or minimal syntactic knowledge for English and Chinese SRL. Although the architectures vary slightly, they are mostly based on a variation of bi-LSTMs. Zhou and Xu (2015); He et al. (2017) connect the layers of LSTM in an interleaving pattern where in (Wang et al., 2015; Marcheggiani et al., 2017) regular bi-LSTM layers are used. Commonly used features for the encoding layer are: pretrained word embeddings; distance from the predicate; predicate context; predicate region mark or flag; POS tag; and predicate lemma embedding. Only a few of the models (Marcheggiani et al., 2017; Marcheggiani and Titov, 2017) perform dependency-based SRL. Furthermore, all methods focus on languages with rich resources and less morphological complexity like English and Chinese.

Character-level Models: Character-level models have proven themselves useful for many NLP tasks such as language modeling (Ling et al., 2015; Kim et al., 2016), POS tagging (Santos and Zadrozny, 2014; Plank et al., 2016), dependency parsing (Dozat et al., 2017) and machine translation (Lee et al., 2017). However the number of comparative studies that analyze their relation to morphology are rather limited. Recently, Vania and Lopez (2017) presented a unified framework, where they investigated the performances of different subword units, namely characters, morphemes and morphological analysis on language modeling task. They experimented with languages of varying morphological typologies and concluded that the performance of character models can not yet match the morphological models, albeit very close. Similarly, Belinkov et al. (2017) analyzed how different word representations help learn better morphology and model rare words on a neural MT task and concluded that character-based representations are much better for learning

morphology.

3 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 $\mathcal{L} = \{roles \cup nonrole\}$, where $roles$ are language-specific semantic roles (mostly consistent with PropBank) and $nonrole$ is a symbol to present tokens that are not arguments. Given θ 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:

$$\hat{\theta} = \arg\min_{\theta} \left(-\sum_{t=1}^{n} log(p(g_t|\theta, s, p)) \right) \quad (1)$$

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

$$\rho(w) = s_0, s_1, .., s_n \quad (2)$$

As proposed by Ling et al. (2015), we treat words as a sequence of subword units. Then, the sequence is fed to a simple bi-LSTM network (Graves and Schmidhuber, 2005; Gers et al., 2000) 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. 4.

$$hs_f, hs_b = \text{bi-LSTM}(s_0, s_1, .., s_n) \quad (3)$$

$$\vec{w} = W_f \cdot hs_f + W_b \cdot hs_b + b \quad (4)$$

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.

$$\vec{x}_t = [\vec{w}; pf_t] \quad (5)$$

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

$$\vec{h_f}, \vec{h_b} = \text{bi-LSTM}(x_t) \quad (6)$$

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

$$p(l_t|\vec{s}, p) = softmax(W_l \cdot [\vec{h_f}; \vec{h_b}] + \vec{b_l}) \quad (7)$$

For simplicity, we assign the label with the highest probability to the input token. [1].

3.1 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 ρ functions. Here, *char* function

ρ	word	output
char	available	$<$-a-v-a-i-l-a-b-l-e-$>$
char3	available	$<$av-ava-vai-ail-ila-lab-abl-ble-le$>$
morph-DEU	prächtiger	[*prächtig*;Pos;Nom;Sg;Masc]
morph-SPA	las	[*el*;postype=article;gen=f;num=p]
morph-CAT	la	[*el*;postype=article;gen=f;num=s]
morph-TUR	boyundaki	[*boy*;NOUN;A3sg;P3sg;Loc;DB;ADJ]
morph-FIN	tyhjyyttä	[*tyhjyys*;Case=Par;Number=Sing]
morph-CZE	si	[*se*;SubPOS=7;Num=X;Cas=3]

Table 1: Sample outputs of different ρ functions

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 (Vania and Lopez, 2017), we excluded these units.

4 Experiments

We use the datasets distributed by LDC for Catalan (CAT), Spanish (SPA), German (DEU), Czech (CZE) and English (ENG) (Hajič et al., 2012b,a); and datasets made available by Haverinen et al. (2015); Şahin and Adalı (2017) for Finnish (FIN) and Turkish (TUR) respectively [2]. Datasets are

---

[1]Our implementation can be found at https://github.com/gozdesahin/Subword_Semantic_Role_Labeling

[2]Turkish PropBank is based on previous efforts (Atalay et al., 2003; Sulubacak et al., 2016; Sulubacak and Eryiğit, 2018; Oflazer et al., 2003; Şahin, 2016b,a)

	#sent	#token	#pred	#role	type
CZE	39K	653K	414K	51	F
ENG	39K	958K	179K	38	F
DEU	36K	649K	17K	9	F
SPA	14K	419K	44K	34	F
CAT	13K	384K	37K	35	F
FIN	12K	163K	27K	20	A
TUR	4K	39K	8K	26	A

Table 2: Training data statistics. A: Agglutinative, F: Fusional

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 Hajič et al. (2009); Haverinen et al. (2015); Şahin and Adalı (2017).

4.1 Experimental Setup

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

Spanish, Catalan: 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.

Finnish: Original dataset defines its own format of semantic annotation, such as 17:PBArgM_mod|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.

Turkish: 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 ρ function to split words into subwords.

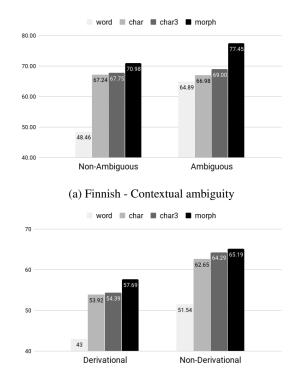**Training and Evaluation:** We lowercase all tokens beforehand and place special start and end of the token characters. For all experiments, we initialized weight parameters orthogonally and used one layer bi-LSTMs both for subword composition and argument labeling with hidden size of 200. Subword embedding 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 [3]. Therefore, we believe omitting PSD helps us gain more useful insights on character level models.

5 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.

---

[3] For instance in English CoNLL-09 dataset, 87% of the predicates are annotated with their first sense, hence even a dummy classifier would achieve 87% accuracy. The best system from CoNLL-09 shared task reports 85.63 F1 on English evaluation dataset, however when the results of PSD are discarded, it drops down to 81.

(a) Finnish - Contextual ambiguity

(b) Turkish - Derivational morphology

Figure 1: Differences in model performances on agglutinative languages

	word	char		char3		morph		
	F1	F1	IOW%	F1	IOW%	F1	IOW%	IOC%
FIN	48.91 51.65	67.24 66.82	37.46	67.78 67.08	38.58	**71.15** **71.88**	45.47	4.97
TUR	44.82 43.14	55.89 54.48	24.68	56.60 55.41	26.28	**59.38** **58.91**	32.48	4.91
SPA	64.30 64.53	67.90 67.64	5.61	68.43 67.64	6.42	**69.39** **69.17**	7.92	2.25
CAT	65.45 65.67	70.56 70.43	7.82	71.34 70.48	9.00	**73.24** **72.36**	11.90	2.66
CZE	63.58 72.69	74.04 74.58	16.45	74.98 75.59	17.93	**80.66** **81.06**	26.87	7.58
DEU	54.78 53.76	63.71 62.75	16.29	65.56 63.70	19.68	**69.35** **72.18**	26.58	5.77
ENG	81.19 78.67	**81.61** **79.22**	0.52	80.65 78.85	-0.67	- -	-	-

Table 3: F1 scores of word, character, character trigram and morphology models for argument labeling. Best F1 for each language is shown in **bold**. First row: results on test, Second row: results on development.

Agglutinative languages have many morphemes attached to a word like beads on a string. This leads to high number of OOV words and cause word lookup models to fail. Hence, the highest *IOW*s by character models are achieved on these languages: Finnish and Turkish. This language family has one-to-one morpheme to meaning mapping with small orthographic differences *(e.g., mış, miş, muş, müş for past perfect tense)*, that can be easily extracted from the data. Even though each morpheme has only one interpretation, each word (consisting of many morphemes) has usually more than one. For instance two possible analyses for the Turkish word "dolar" are (1) "dol+Verb+Positive+Aorist+3sg" *(it fills)*, (2) "dola+Verb+Positive+Aorist+3sg" *(he/she wraps)*. For a syntactic task, models are not obliged to learn the difference between the two; whereas for a semantic task like SRL, they are. We will refer to this issue as *contextual ambiguity*. Another important linguistic issue for agglutinative languages is the complex interaction between morphology and syntax, which is usually achieved via derivational morphemes. In other words, unlike *inflectional* morphemes that only give information on *word-level semantics*, derivational morphemes provide more clues on *sentence-level semantics*. The effects of these two phenomena on model performances is shown in Fig. 1. Scores given in Fig. 1 are absolute F1 scores for each model. For the analysis in Fig. 1a, we separately calculated F1 scores of each model on words that have been observed with at least two different set of morphological features (*ambiguous*), and one set of features (*non-ambiguous*). Due to the low number of ambiguous words in Turkish dataset ($\leq$100), it has been calculated for Finnish only. Similarly, for the derivational morphology analysis in Fig. 1b, we have separately calculated scores for sentences containing derived words (*derivational*), and simple sentences without any derivations. Both analyses show that access to gold morphological tags (*oracle*) provided big performance gains on arguments with contextual ambiguity and sentences with derived words. Moderate *IOC* signals that *char* and *char3* learns to imitate the "beads" and their "predictable order" on the string (in the absence of the aforementioned issues).
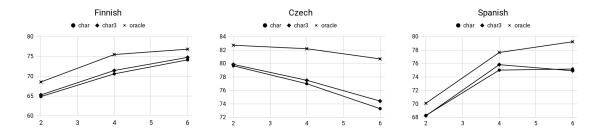
Figure 2: *x axis*: Number of morphological features; *y axis*: Targeted F1 scores

Fusional languages *may* have many morphemes in a word. Spanish and Catalan have relatively low morpheme per word ratio that results with low OOV% (5.63 and 5.40 for Spanish and Catalan respectively); whereas, German and Czech have OOV% of 7.93 and 7.98 (Hajič et al., 2009). We observe that *IOW* by character models are well aligned with OOV percentages of the datasets. Unlike agglutinative languages, single morpheme can serve multiple purposes in fusional languages. For instance, "o" (e.g., *habl-o*) may signal 1_{st} person singular present tense, or 3_{rd} person singular past tense. We count the number of surface forms with at least two different features and use their ratio *(#ambiguous forms/#total forms)* as a proxy to morphological complexity of the language. The *complexities* are approximated as 22%, 16%, 15% for Czech, Spanish and Catalan respectively; which are aligned with the observed *IOC*s. Since there is no unique morpheme to meaning mapping, generally multiple morphological tags are used to resolve the *morpheme ambiguity*. Therefore there is an indirect relation between the number of morphological tags used and the ambiguity of the word. To demonstrate this phenomena, we calculate targeted F1 scores on arguments with varying number of morphological features. Results using feature bins of [1-2], [3-4] and [5-6] are given in Fig. 2. As the number of features increase, the performance gap between oracle and character models grows dramatically for Czech and Spanish, while it stays almost fixed for Finnish. This finding suggests that high number of morphological tags signal the vagueness/complex cases in fusional languages where character models struggle; and also shows that the complexity can not be directly explained by number of morphological tags for agglutinative languages. German is known for having many compound words and compound lemmas that lead to high OOV% for lemma; and also is less ambiguous (9%). Therefore we would expect a lower *IOC*. However, the evaluation set consists only of 550 predicates and 1073 arguments, hence small changes in prediction lead to dramatic percentage changes.

5.1 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:

$$p_{final} = f(p_0, p_1, .., p_n | \phi) \qquad (8)$$

where f is the combining function with parameter ϕ. The simplest global approach is *averaging (AVG)*, where f is simply the mean function and p_is 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 ϕ 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)*. [4]

Although not guaranteed, diverse models can be achieved by altering the input representation,

---

[4]To train the SG model, we have used one linear layer with 64 hidden units followed by sigmoid nonlinear activation. Weights are orthogonally initialized and optimized via adam algorithm with a learning rate of 0.02 for 25 epochs.

	char+char3			char+oracle			char3+oracle		
	Avg	SG	*IOB%*	Avg	SG	*IOB%*	Avg	SG	*IOB%*
Czech	76.24	76.26	**2.03**	80.36	81.06	*0.49*	80.57	81.10	*0.55*
Finnish	70.31	70.29	**4.58**	72.73	72.88	*2.42*	72.72	73.02	*2.62*
Turkish	59.43	59.39	**6.34**	61.98	62.07	*4.53*	60.56	60.74	*2.28*
Spanish	70.01	70.05	*3.16*	71.80	71.75	**3.47**	71.64	71.62	**3.24**
Catalan	72.79	72.71	*2.03*	74.80	74.82	**2.16**	75.15	75.18	**2.66**
German	66.84	66.97	*2.15*	71.02	71.16	**2.62**	71.31	71.25	**2.84**

Table 4: Results of ensembling via averaging (Avg) and stack generalization (SG). *IOB: Improvement Over Best of baseline models*

the learning algorithm, training data or the hyper-parameters. 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.

6 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.

6.1 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.

6.2 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*.

6.3 Out-of-Domain (OOD) Data

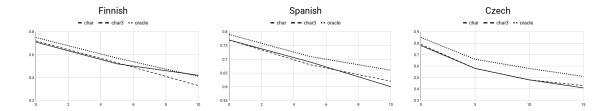As part of the CoNLL09 shared task (Hajič et al., 2009), out of domain test sets are provided for

Figure 3: *X axis*: Distance between the predicate and the argument, *Y axis*: F1 scores on argument labels
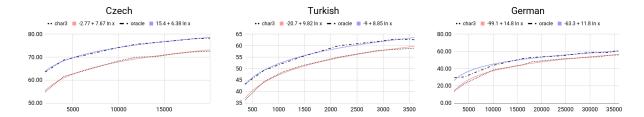

Figure 4: Performance of units w.r.t training data size. *X axis*: Number of sentences, *Y axis*: F1 score

	word	char	*IOW%*	char3	*IOW%*	oracle	*IOW%*	*IOC%*
CZE	69.97	72.98	*4.30*	**73.24**	*4.67*	72.28	*3.30*	*-1.31*
DEU	51.50	**57.05**	*10.78*	55.75	*8.24*	38.51	*-25.24*	*-45.17*
ENG	66.47	68.83	*0.70*	**70.22**	*0.23*	-	-	-

Table 5: F1 scores on out of domain data. Best scores are shown with **bold**.

		char3		oracle	
		F1	*I (%)*	F1	*I (%)*
Finnish	$\ell = 1$	67.78		71.15	
	$\ell = 2$	67.62	*-0.2*	75.71	*6.4*
Turkish	$\ell = 1$	56.60		59.38	
	$\ell = 2$	56.93	*0.5*	61.02	*2.7*
Spanish	$\ell = 1$	68.43		69.39	
	$\ell = 2$	69.30	*1.3*	71.56	*3.1*
Catalan	$\ell = 1$	71.34		73.24	
	$\ell = 2$	71.71	*0.5*	74.84	*2.2*

Table 6: Effect of layer size on model performances. *I*: Improvement over model with one layer.

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 (Hajič et al., 2009); 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.

6.4 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.

6.5 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
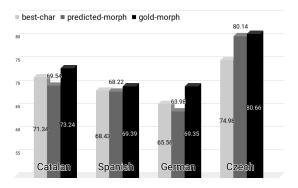
Figure 5: F1 scores for *best-char* (best of the CLMs) and model with predicted (*predicted-morph*) and gold morphological tags (*gold-morph*).

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.

7 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:

- For in-domain data, character-level models cannot yet match the performance of morphology-level models. However, they still provide considerable advantages over whole-word models,

- Their shortcomings depend on the morphology type. For agglutinative languages, their performance is limited on data with rich *derivational morphology* and high *contextual ambiguity* (morphological disambiguation); and for fusional languages, they struggle on tokens with high number of morphological tags,

- Similarity between character and morphology-level models is higher than the similarity within character-level (char and char-trigram) models on languages with high OOV%; and vice versa,

- Their ability to handle long-range dependencies is very similar to morphology-level models,

- They rely relatively more on training data size. Therefore, given more training data their performance will improve faster than morphology-level models,

- They perform *substantially* well on out of domain data, surpassing all morphology-level models. However, relatively less improvement is expected when model complexity is increased,

- They generally perform better than models that only have access to predicted/silver morphological tags.

8 Acknowledgements

References

Nart Bedin Atalay, Kemal Oflazer, and Bilge Say. 2003. The Annotation Process in the Turkish Treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora, LINC at EACL 2003, Budapest, Hungary, April 13-14, 2003*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 861–872.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research* 12:2461–2505.

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* pages 20–30.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic Role Labeling with Neural Network Factors. In *EMNLP*. pages 960–970.

Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* 12(10):2451–2471.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5-6):602–610.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Stroudsburg, PA, USA, CoNLL '09, pages 1–18.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Adam Meyers, Jan Štěpánek, Joakim Nivre, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2012a. 2009 CoNLL Shared Task Part 1 LDC2012T04. Web Download.

Jan Hajič, Maria A. Martí, Lluis Marquez, Joakim Nivre, Jan Štěpánek, Sebastian Padó, and Pavel Straňák. 2012b. 2009 CoNLL Shared Task Part 1 LDC2012T03. Web Download.

Katri Haverinen, Jenna Kanerva, Samuel Kohonen, Anna Missila, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. The Finnish Proposition Bank. *Language Resources and Evaluation* 49(4):907–926.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-Aware Neural Language Models. In *AAAI*. pages 2741–2749.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *TACL* 5:365–378.

Wang Ling, Tiago Luis, Luis Marujo, Ramon F Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *EMNLP*. pages 1520–1530.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A Simple and Accurate Syntax-Agnostic Neural Model for Dependency-based Semantic Role Labeling. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 411–420.

Diego Marcheggiani and Ivan Titov. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1507–1516.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In *Treebanks*, Springer, pages 261–277.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Gözde Gül Şahin and Eşref Adalı. 2017. Annotation of semantic roles for the Turkish Proposition Bank. *Language Resources and Evaluation* pages 1–34.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pages 1818–1826.

Umut Sulubacak and Gülşen Eryiğit. 2018. Implementing Universal Dependency, Morphology and Multiword Expression Annotation Standards for Turkish Language Processing. *Turkish Journal of Electrical Engineering Computer Sciences* pages 1–23.

Umut Sulubacak, Tuğba Pamay, and Gülşen Eryiğit. 2016. IMST: A Revisited Turkish Dependency Treebank. In *Proceedings of the 1st International Conference on Turkic Computational Linguistics (TurCLing) at CICLing, Konya, Turkey, 2016*.

Clara Vania and Adam Lopez. 2017. From Characters to Words to in Between: Do We Capture Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 2016–2027.

Zhen Wang, Tingsong Jiang, Baobao Chang, and Zhi-fang Sui. 2015. Chinese Semantic Role Labeling with Bidirectional Recurrent Neural Networks. In *EMNLP*. pages 1626–1631.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. pages 1127–1137.

Gözde Gül Şahin. 2016a. Framing of Verbs for Turkish PropBank. In *In Proceedings of 1st International Conference on Turkic Computational Linguistics, TurCLing*.

Gözde Gül Şahin. 2016b. Verb Sense Annotation for Turkish PropBank via Crowdsourcing. In *Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016, Konya, Turkey, April 3-9, 2016, Revised Selected Papers, Part I*. pages 496–506.