Character-Level Models (CLMs) vs Morphology in Semantic Role Labeling

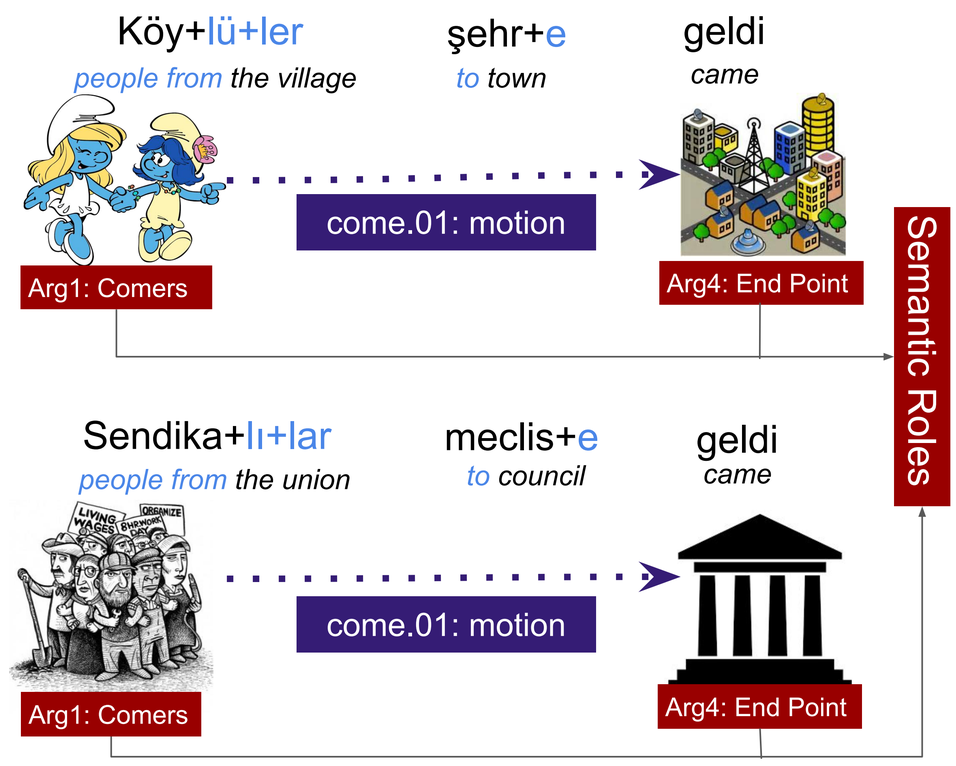
Gözde Gül Şahin*, Mark Steedman

goezde.guel@gmail.com, steedman@inf.ed.ac.uk

Ubiquitous Knowledge Processing (UKP) Lab, University of Edinburgh

*work done partiallly in Istanbul Technical University (İTÜ) and University of Edinburgh

The Role of Morphology in SRL

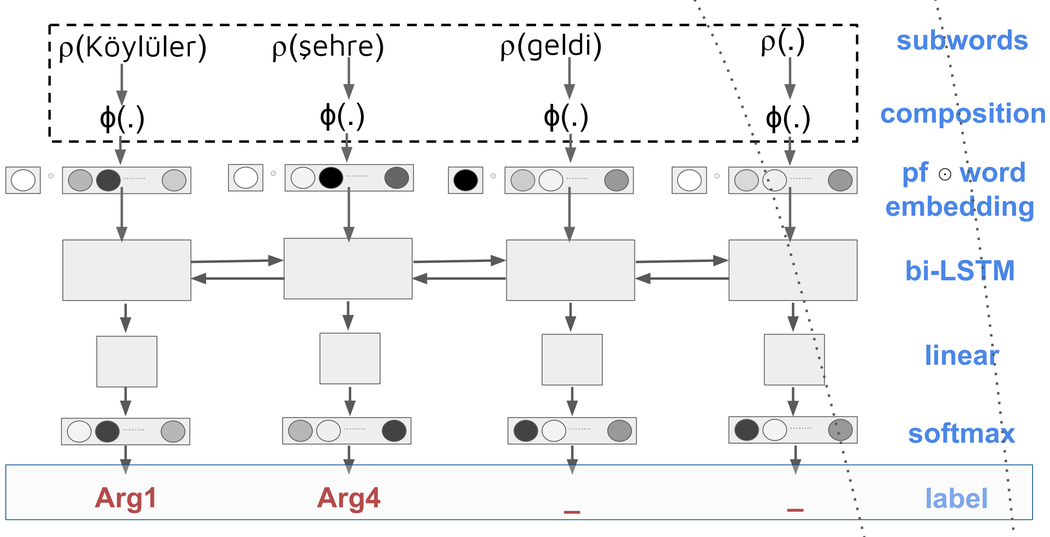


Morphology is essential for **semantic role labeling** (**SRL**) but *expensive*, so we ask the following questions:

Can **character level models (CLMs)** replace **oracle** (gold morphological analysis)?

What are CLMs' **Limitations** and **Strengths** compared to oracle ?

Method



1 Words are decomposed into subword units

Decomposition	ρ(Köylüler)
word	köylüler
char	<-k-ö-y-l-ü-l-e-r->
char3	<kö-köy-öyl-ylü-lül-üle-ler-er>
morph	köylü-Noun-NAdj-A3pl-Pnon-Nom

2 Subwords are composed into word vectors. Words are treated as a sequence of subwords.

hs_f, hs_b = bi-LSTM(<-k-ö-y-l-ü-l-e-r ->)

$w = W_f . hs_f + W_b . hs_b + b$

3 Predicate flag (pf) is concatenated to *w*.

4 Generated input vectors are fed to the sequence-labeling network. For the sake of simplicity, the label with the highest probability is assigned to the input.

Experiments

Dataset:

CoNLL-09 dependency-based SRL **shared task dataset** for **Czech**, **Spanish**, **Catalan**, **German** and **English** and **Free PropBanks**: **Turkish** Propbank [3] and **Finnish** PropBank

	word	char		char-trigram		gold morph. tags		
	F1	F1	IOW%	F1	IOW%	F1	IOW%	IOC%
FINNISH	48.91	67.24	37.46	67.78	38.58	**71.15**	45.47	4.97
TURKISH	44.82	55.89	24.68	56.60	26.28	**59.38**	32.48	4.91
SPANISH	64.30	67.90	5.61	68.43	6.42	**69.39**	7.92	2.25
CATALAN	65.45	70.56	7.82	71.34	9.00	**73.24**	11.90	2.66
CZECH	63.58	74.04	16.45	74.98	17.93	**80.66**	26.87	7.58
GERMAN	54.78	63.71	16.29	65.56	19.68	**69.35**	26.58	5.77
ENGLISH	81.19	**81.61**	0.52	80.65	-0.67	-	-	-

Table 1: Argument labeling F1 scores for each subword unit and language.*

The best model was the **morphology-level model** in all languages, BUT...

Why does Improvement over Word (**IOW**) range between **0%-38% ??**

Why does Improvement over Character (**IOC**) range between **2%-10%?**

*These are the results on test set. Please see the paper for development data results.

Explaining Improvement over Word *is easy*

IOW rates are well aligned with Out-of-Vocabulary (OOV)%

Highest IOW in agglutinative languages ← Many morphemes attached to a word *(e.g.,belge-len-dir-il-eme-yen-ler)*
Moderate IOW in German, Czech ← ~7.9% of the test tokens are not seen in the training
Low IOW in Spanish, Catalan ← ~5% of the test tokens are not seen in the training

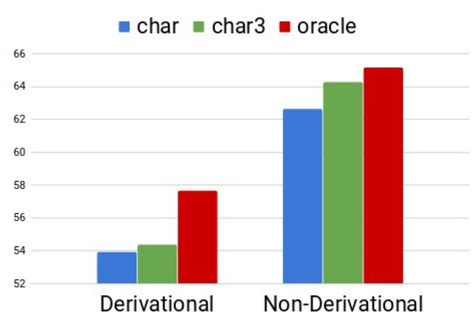
Explaining Improvement over Character *is hard!*

...Same as answering when MLM provides more structure then CLM

Agglutinative

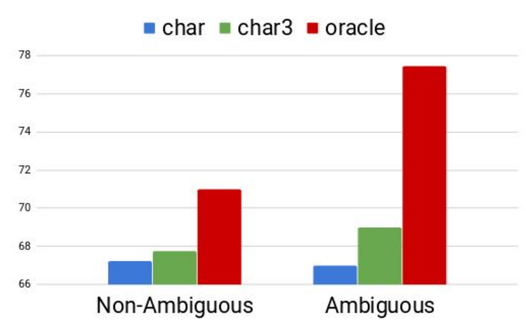
1 Derivational morphology

Yaz+Iık | ev+iniz+de ^DB ydi
Summer house+your+at ^DB it+was



2 Contextual Ambiguity

(1) **dol**+Verb+Positive+Aorist+3sg "It **fills**"
(2) **dola**+Verb+Positive+Aorist+3sg "He **wraps**"

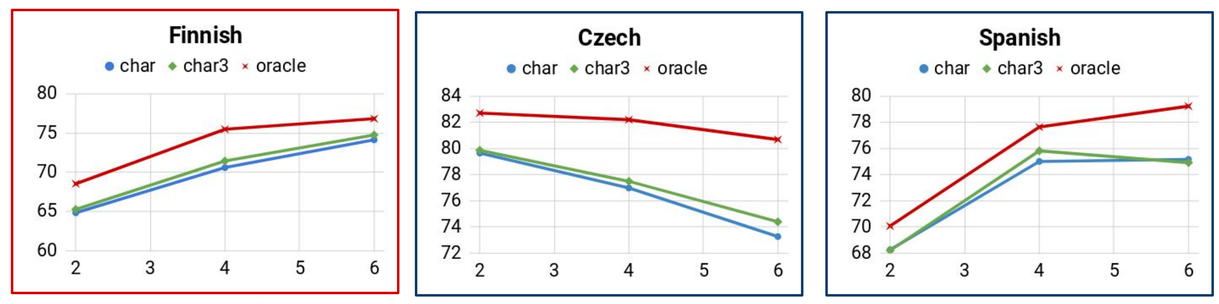


Fusional

Morpheme ambiguity

morpheme to meaning mapping is 1:many

rule | perfective | 2nd person | singular
re:k | s | is | ti

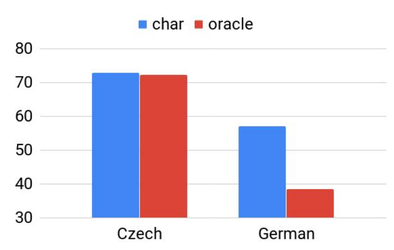


*Red frame: Agglutinative, Blue: Fusional

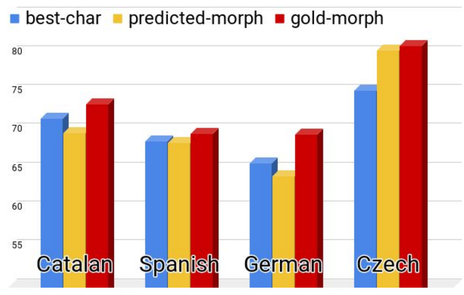
X: Number of morphological tags, Y: F1 score

When are CLMs better/worse than MLMs ?

On Out-of-Domain Data



When only predicted tags are available



Large training data

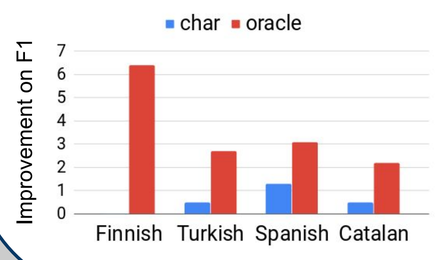
F1 growth is logarithmic w.r.t datasize (x)

Curve Equations for Czech

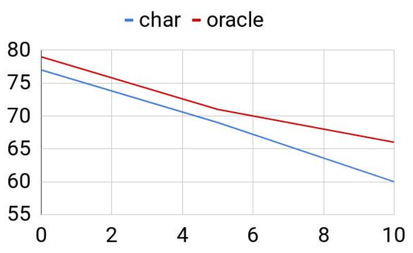
char: -2.77 + **7.67** ln x
oracle: 15.4 + **6.38** ln x

x coefficients are higher for char in all languages

Adding More Layers



Long-range Dependencies



Small training data

char: **-2.77** + 7.67 ln x
oracle: **15.4** + 6.38 ln x

Initial value is higher for oracle in all languages

So the answers are...

- For in-domain data, CLMs can not yet match the performance of MLMs, but surpass WLMs by a large margin
- Its shortcomings depend on the language type. The hard cases are: Derivational morphology and contextual ambiguity for agglutinative languages; and tokens with many morphological tags in fusional languages.
- They perform better on out-of-domain data; when there is only access to predicted tags; and when a large enough training set is available. Targeted scores for long range dependencies are similar.
- They don't benefit as much from increasing of the model size and perform worse in case of small training data size.

Acknowledgments

Gözde Gül Şahin 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)** in Fall 2015. This work was supported by **ERC H2020 Advanced Fellowship GA 742137 SEMANTAX** and a **Google Faculty award** to Mark Steedman.