Automatic Morpheme Segmentation and Labeling in Universal Dependencies resources

Miikka Silfverberg
Mans Hulden

May 22, 2017

UD & Morphosyntactic info

- UD resources for many languages contain rich morphological labeling for lexical and grammatical properties of words

Lexical features	Inflectional features	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
Foreign	Definite	Voice
Abbr	Degree	Evident
		Polarity
		Person
		Polite

http://universaldependencies.org/u/feat/index.html

Segmentation and labeling

English Data

word form	lemma	POS		morphosyntactic features
busted	bust	VERB	VBN	Tense=Past|VerbForm=Part
authorities	authority	NOUN	NNS	Number=Plur
announced	announce	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin
cells	cell	NOUN	NNS	Number=Plur
cell	cell	NOUN	NN	Number=Sing
mid-nineties	mid-ninety	NOUN	NNS	Number=Plur
tension	tension	NOUN	NN	Number=Sing
tensions	tension	NOUN	NNS	Number=Plur
announcing	announce	VERB	VBG	Tense=Pres|VerbForm=Part
authority	authority	NOUN	NN	Number=Sing
killed	kill	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin

Segmentation and labeling

English Data

busted	bust	VERB	VBN	Tense=Past|VerbForm=Part
authorities	authority	NOUN	NNS	Number=Plur
announced	announce	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin
cells	cell	NOUN	NNS	Number=Plur
cell	cell	NOUN	NN	Number=Sing
mid-nineties	mid-ninety	NOUN	NNS	Number=Plur
tension	tension	NOUN	NN	Number=Sing
tensions	tension	NOUN	NNS	Number=Plur
announcing	announce	VERB	VBG	Tense=Pres|VerbForm=Part
authority	authority	NOUN	NN	Number=Sing
killed	kill	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin

Segmentation and labeling

English Data

busted	bust	VERB	VBN	Tense=Past|VerbForm=Part
authorities	authority	NOUN	NNS	Number=Plur
announced	announce	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin
cells	cell	NOUN	NNS	Number=Plur
cell	cell	NOUN	NN	Number=Sing
mid-nineties	mid-ninety	NOUN	NNS	Number=Plur
tension	tension	NOUN	NN	Number=Sing
tensions	tension	NOUN	NNS	Number=Plur
announcing	announce	VERB	VBG	Tense=Pres|VerbForm=Part
authority	authority	NOUN	NN	Number=Sing
killed	kill	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin

Segmentation and labeling

English Data

busted	bust	VERB	VBN	Tense=Past|VerbForm=Part
authorities	authority	NOUN	NNS	Number=Plur
announced	announce	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin
cells	cell	NOUN	NNS	Number=Plur
cell	cell	NOUN	NN	Number=Sing
mid-nineties	mid-ninety	NOUN	NNS	Number=Plur
tension	tension	NOUN	NN	Number=Sing
tensions	tension	NOUN	NNS	Number=Plur
announcing	announce	VERB	VBG	Tense=Pres|VerbForm=Part
authority	authority	NOUN	NN	Number=Sing
killed	kill	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin

Could labeling/alignments be performed automatically?

Linguistic intuitions in segmentation

Data

```
announced
announcing
authorities
busted
cell
cells
killed
mid-nineties
```

Allomorphs

Number=Plur

Tense=Past

Number=Sing

(Lemmas)

Raw data

Linguistic intuitions in segmentation

Data

```
announ ced
annou ncing
authorit ies
buste d
ce ll
cel ls
kil led
mid-nine ties
```

Bad Segmentation

Allomorphs

Number=Plur

```
ies
ties
ls
```

(Stems)

```
announ
annou
authorit
buste
ce
cel
kil
mid-nine
```

Tense=Past

```
ced
d
led
```

Number=Sing

```
ll
```

VerbForm=Part

```
ncing
```

Linguistic intuitions in segmentation

Data

```
announ ced
annou ncing
authorit ies
buste d
ce ll
cel ls
kil led
mid-nine ties
```

Bad Segmentation

Allomorphs

Number=Plur

```
ies
ties
ls
```

(Stems)

```
announ
annou
authorit
buste
ce
cel
kil
mid-nine
```

Tense=Past

```
ced
d
led
```

Number=Sing

```
ll
```

VerbForm=Part

```
ncing
```

Linguistic intuitions in segmentation

Data

```
announ ced
annou ncing
authorit ies
buste d
ce ll
cel ls
kil led
mid-nine ties
```

Bad Segmentation

Allomorphs

Number=Plur

```
ies
ties
ls
```

(Stems)

```
announ
annou
authorit
buste
ce
cel
kil
mid-nine
```

Tense=Past

```
ced
d
led
```

Number=Sing

```
ll
```

VerbForm=Part

```
ncing
```

Linguistic intuitions in segmentation

Data

announc ed
announc ing
authoriti es
bust ed
cell ∅
cell s
kill ed
mid-nineti es

Better Segmentation

Allomorphs

Number=Plur

s
es

Tense=Past

ed

Number=Sing

∅

(Stems)

announc
authoriti
bus
cell
kill
mid-nineti

VerbForm=Part

ing

Linguistic intuitions in segmentation

Data

announc ed
announc ing
authoriti es
bust ed
cell ∅
cell s
kill ed
mid-nineti es

Allomorphs

Number=Plur
s
es

Tense=Past
ed

Number=Sing
∅

(Stems)
announc
authoriti
bus
cell
kill
mid-nineti

VerbForm=Part
ing

Better Segmentation

Linguistic intuitions in segmentation

Data

announc ed
announc ing
authoriti es
bust ed
cell ∅
cell s
kill ed
mid-nineti es

Allomorphs

Number=Plur

s
es

Tense=Past

ed

Number=Sing

∅

(Stems)

announc
authoriti
bus
cell
kill
mid-nineti

VerbForm=Part

ing

Better Segmentation

Interim

- This is an inference problem with weak supervision
- We can treat this as a search problem in the (huge!) space of all possible segmentations and labelings over a (UD) corpus
- The labels given by UD give us a weak supervision (we know which features are present in each word form, and which aren't)

```
cells        cell        NOUN   NNS  Number=Plur
cell         cell        NOUN   NN   Number=Sing
```

Objective function (I)

- A simple objective would be to minimize the total number of allomorph types in the corpus

Minimize allomorphs

Data

announc ed
announc ing
authoriti es
bust ed
cell ∅
cell s
kill ed
mid-nineti es

Better Segmentation

Allomorphs

Number=Plur

s
es

(Lemmas)

announc
authoriti
bus
cell
kill
mid-nineti

Tense=Past

ed

Number=Sing

∅

VerbForm=Part

ing

allomorphs=11

Minimize allomorphs

Data

```
announ ced
annou ncing
authorit ies
buste d
ce ll
cel ls
kil led
mid-nine ties
```

Bad Segmentation

Allomorphs

Number=Plur

```
ies
ties
ls
```

(Lemmas)

```
announ
annou
authorit
buste
ce
cel
kil
mid-nine
```

Tense=Past

```
ced
d
led
```

Number=Sing

```
ll
```

VerbForm=Part

```
ncing
```

allomorphs=16

Objective function (I)

- A simple objective would be to minimize the total number of allomorph types in the corpus
- Unfortunately, that function has less desirable properties (not continuous, not differentiable, insensitive to small changes in segmentation)

```
kil led
bus ted
announ ced
talk ed
cal led
```

Objective function (I)

- A simple objective would be to minimize the total number of allomorph types in the corpus
- Unfortunately, that function has less desirable properties (not continuous, not differentiable, insensitive to small changes in segmentation)

```
kill ed
bus ted
announ ced
talk ed
cal led
```

Objective function (I)

- A simple objective would be to minimize the total number of allomorph types in the corpus
- Unfortunately, that function has less desirable properties (not continuous, not differentiable, insensitive to small changes in segmentation)

```
kill ed
bus ted
announ ced
talk ed
cal led
```

No change in allomorph count!

Objective function (II)

"Symmetric conditional probability"

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

s is segment (substring)
f is feature-value pair

all substrings

set of labels

Intuitively: (1) substrings declared allomorphs should be reliable predictors of a feature; (2) features should predict a substring

Encourages few different features per allomorph, and few allomorphs per feature

Objective function (II)

"Symmetric conditional probability"

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

s is segment (substring)
f is feature-value pair

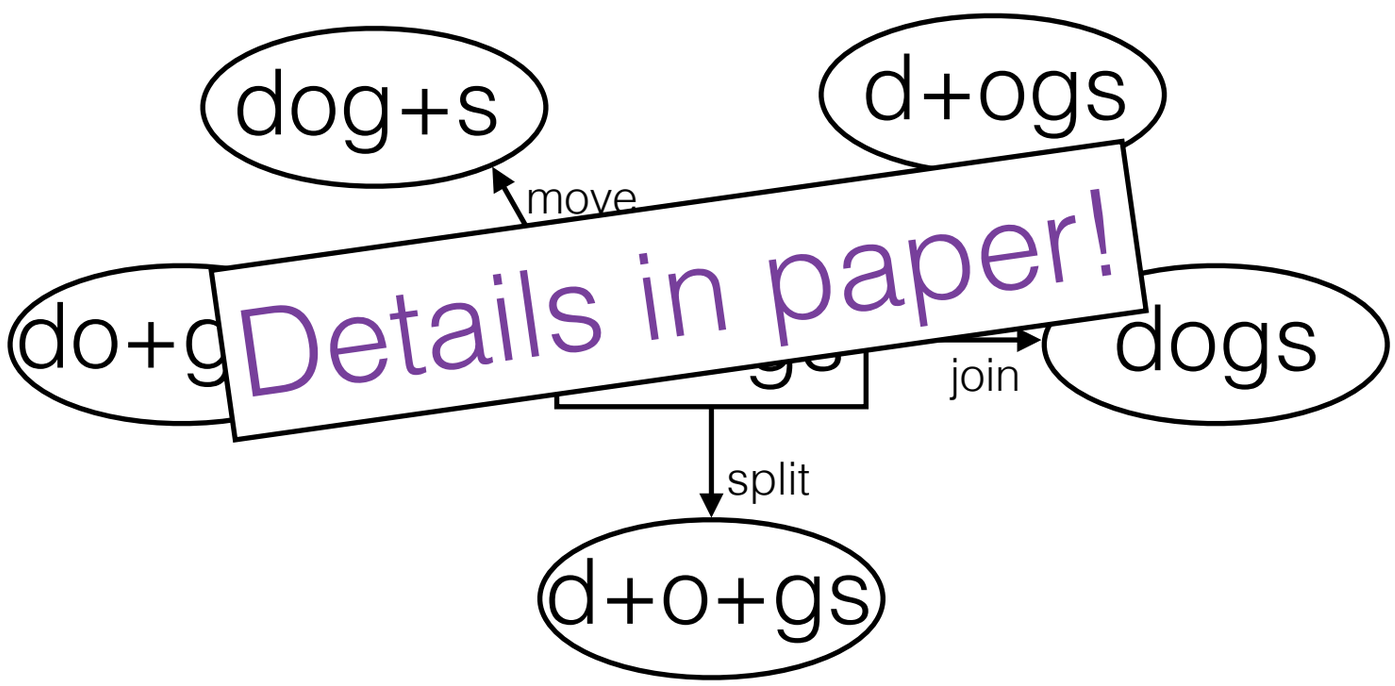
all substrings

set of labels

Not strictly true in either direction, but a workable proxy for the allomorph minimization idea (e.g. English has plurals -s,-es, ∅; and -s can be both pluralizer and present tense 3p allomorph)

Sampling

- Search space is still exponential, so we use a sampling approach
- Start from a random segmentation and labeling
- Make small changes and probabilistically move in the direction that increases the objective function

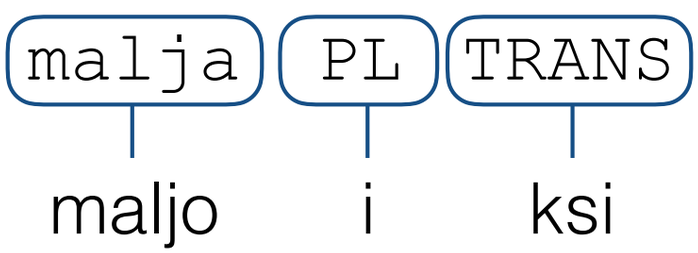


Sampling

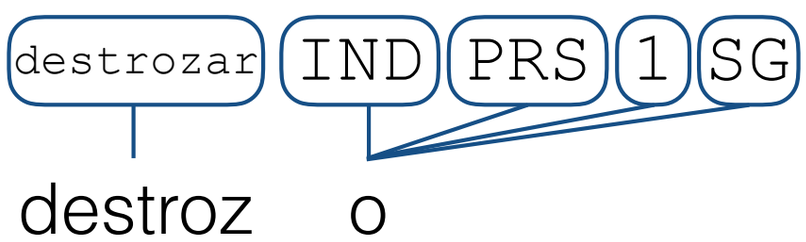
- Allow many features to correspond to single segments
- Also allow features to correspond to a special NULL (∅) segment (similar to IBM alignment models)
- Don't allow overlap in features

forbidden: f_1 f_2 **xyz**

- All actually occurring letters must be associated with a label



Finnish (agglutinative)



Spanish (fusional)

Evaluation

- We compare the method with an unsupervised segmenter Morfessor (Creutz and Lagus, 2005)
- Run Morfessor to get a segmentation, then assign labels to maximize objective function with given segmentations (a much easier problem if segmentation is given)
- We use hand-segmented and aligned gold data for Finnish, Swedish, Spanish (a few hundred word forms each) from CoNLL UD shared task this year

Evaluation

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

Morpheme boundaries

Evaluation

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

Unlabeled morphemes

Evaluation

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

Labeled morphemes

Wrapup

- Weak supervision helps in inducing segmentation and allomorph labeling
- Can be run on all UD languages
- Gives a consistent segmentation & labeling
- Code at https://github.com/mpsilfve/ud-segmenter
- Errors (differences to linguist-preferred gold standard) remain, some due to objective function, some probably due to well-known differing linguistic notions about gold standard (is it announc+ed or announce+d?), frequency effects
- Future work: evaluate a range of objective functions, implement raw allomorph minimization

Thank You