Tree models, syntactic functions and word representations

Simon Šuster*, Gertjan van Noord* and Ivan Titov◊

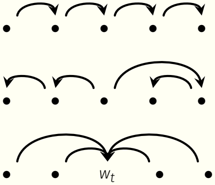
*University of Groningen, ◊University of Amsterdam

Word representations

- Categorical or vectorial object associated with a word
- Way of telling which words are (semantically) similar
- Improve generalization in NLP applications

- Categorical or vectorial object associated with a word
- Way of telling which words are (semantically) similar
- Improve generalization in NLP applications

Definition of context:



+ syntax

Word embeddings
[Levy and Goldberg, 2014]

Clustering
[Šuster and Van Noord, 2014]

Word-space models (DS)
+ dim. red.
[Padó and Lapata, 2007]

Other (probabilistic) models
e.g. HMMs
[Grave et al., 2013]

1. Reproducing [Grave et al., 2013]:
   dependency trees provide better context than sequences
2. Extend tree HMMs with syntactic functions

Model introduction

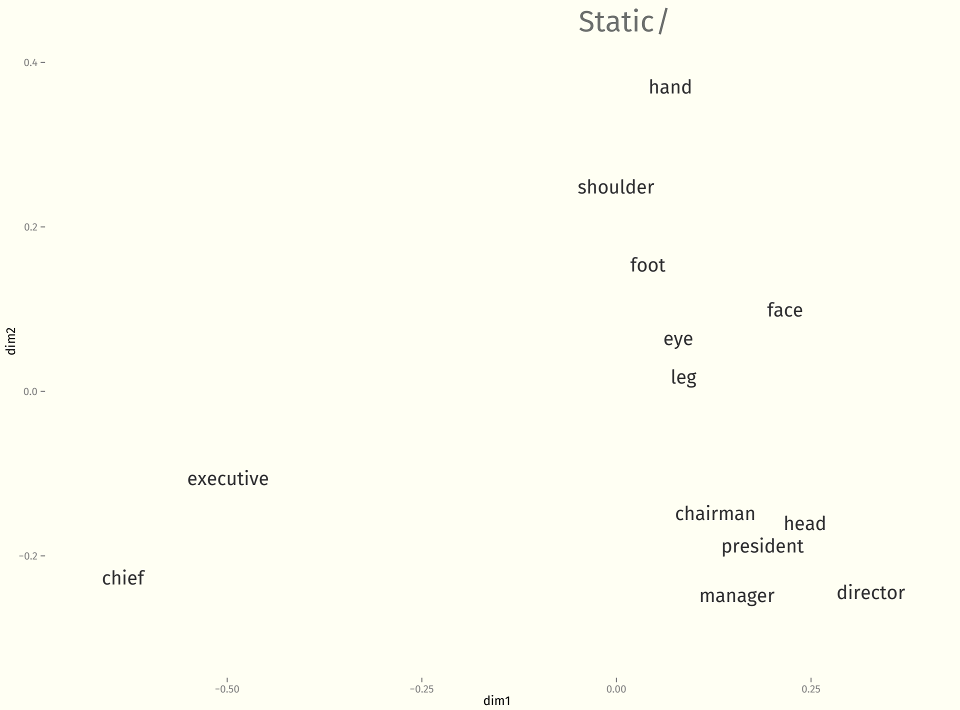
Based on Hidden Markov tree models

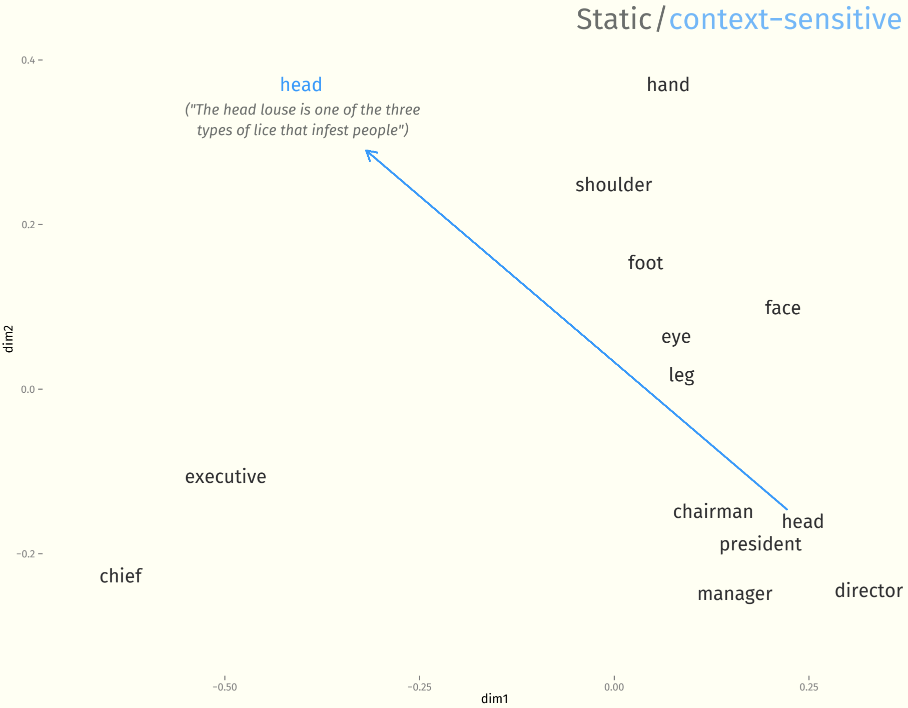
- Word representation from the hidden layer
- Think of state as semantic class
- Number of states set beforehand
- **Context-sensitive** decoding (polysemy)

Decoding

- Categorical: max-product (Viterbi)
  or
- Continuous: state posterior distribution

- Context-sensitive
  or
- Static
  1. average posterior distributions per word type
  2. then use these vectors when needed (context-*in*sensitive)


Static /
hand
shoulder
foot
face
eye
leg
executive
chairman
head
president
chief
manager
director
dim1
dim2
0.4
0.2
0.0
−0.2
−0.50
−0.25
0.00
0.25



Static / context-sensitive
head
("The head louse is one of the three types of lice that infest people")
hand
shoulder
foot
face
eye
leg
executive
chairman
head
president
chief
manager
director
dim1
dim2
0.4
0.2
0.0
-0.2
-0.50
-0.25
0.00
0.25

Training

- Online EM with sum-product message passing
  - state splitting, final 128 states
  - Brown initialization
  - sparse approximate vectors

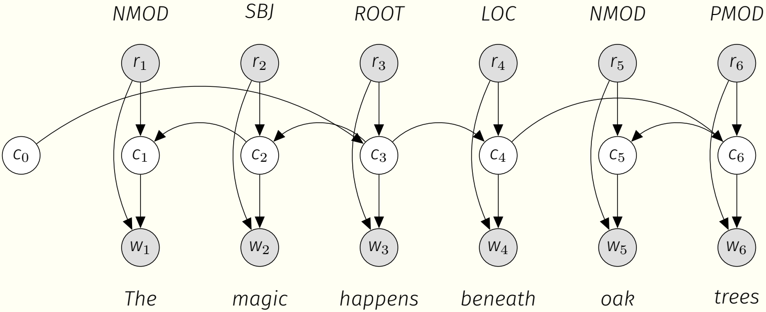
Why syntactic functions

- Prevent sharing same parameters by all children of a node
- Account for (semantically) different children across syntactic functions

Discriminating between types of contexts

Syntactic function: **additional observed variable** in the model

- Modulates transitions and emissions, cf. [Bengio and Frasconi, 1996]

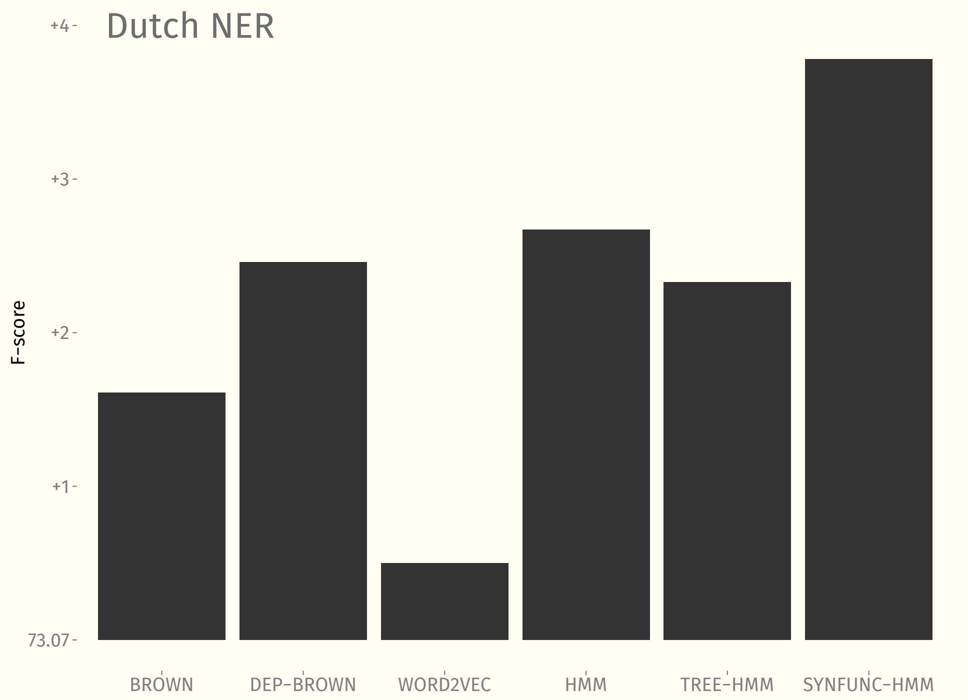


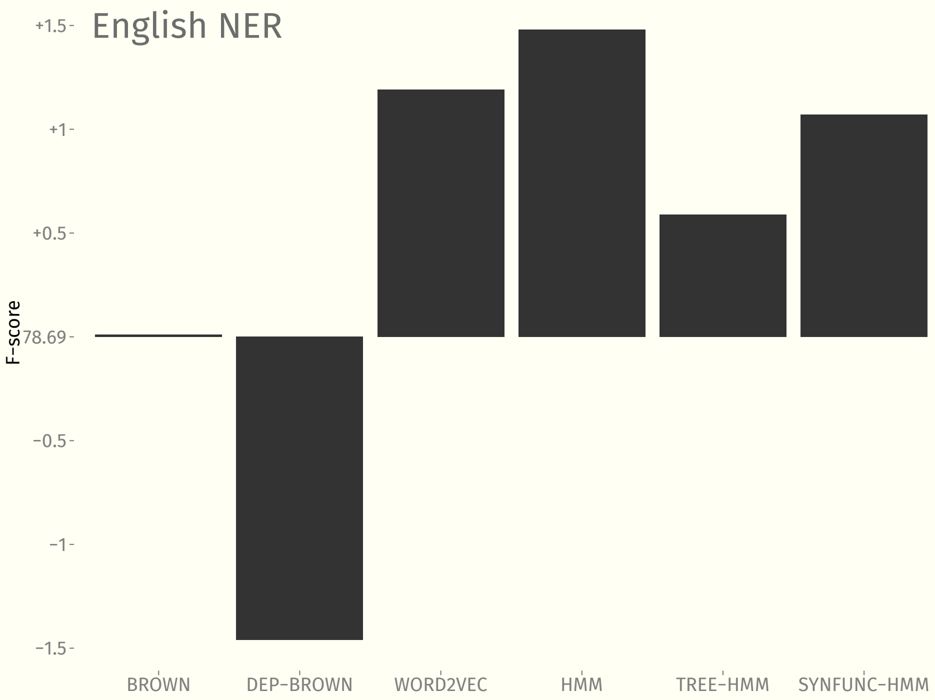
- In practice, can't get reliable estimates for **all** syntactic functions

Evaluate on CoNLL tasks for English and Dutch

Approach

- Structured averaged perceptron
- Several lexical features as baseline [Turian et al., 2010]
- Add word representations (128-dimensional) as features

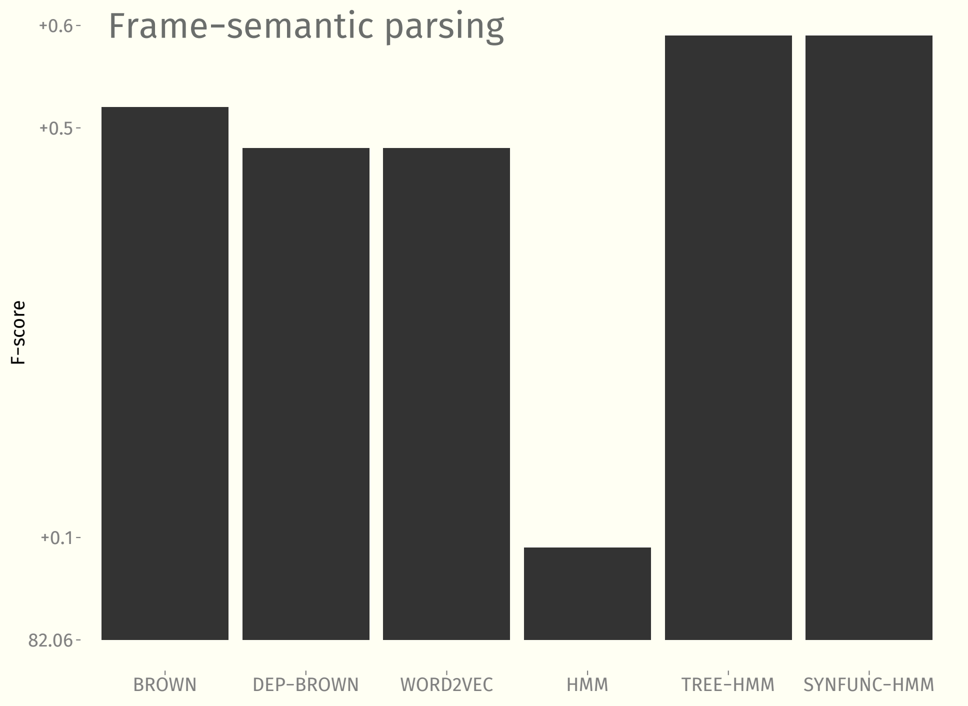

Dutch NER
F−score
+4
+3
+2
+1
73.07
BROWN
DEP−BROWN
WORD2VEC
HMM
TREE−HMM
SYNFUNC−HMM



English NER
F−score
+1.5
+1
+0.5
78.69
−0.5
−1
−1.5
BROWN
DEP−BROWN
WORD2VEC
HMM
TREE−HMM
SYNFUNC−HMM

- Which predicate evokes which frame (**frame identification**)
- Which are the arguments constituting the frame (argument identification)

- Semafor [Das et al., 2014]
- FrameNet


Frame−semantic parsing
F−score
+0.6
+0.5
+0.1
82.06
BROWN
DEP−BROWN
WORD2VEC
HMM
TREE−HMM
SYNFUNC−HMM

Contributions

1. Reproducing [Grave et al., 2013]:
   dependency trees provide better context than sequences
   - not robust
2. Extend tree HMM with syntactic functions
   - works in certain cases

Bengio, Y. and Frasconi, P. (1996).
Input-output HMMs for sequence processing.
IEEE Transactions on Neural Networks, 7(5).

Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2014).
Frame-semantic parsing.
Computational Linguistics, 40(1):9–56.

Grave, E., Obozinski, G., and Bach, F. (2013).
Hidden Markov tree models for semantic class induction.
In *CoNLL*.

Levy, O. and Goldberg, Y. (2014).
Dependency-based word embeddings.
In *ACL*.

Padó, S. and Lapata, M. (2007).
Dependency-based construction of semantic space models.
Computational Linguistics, 33:161–199.

Turian, J., Ratinov, L., and Bengio, Y. (2010).

Word representations: a simple and general method for semi-supervised learning.
In *ACL*.

Šuster, S. and Van Noord, G. (2014).
From neighborhood to parenthood: the advantages of dependency representation over bigrams in Brown clustering.
In *COLING*.