Sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian

Bálint Döbrössy[1] Márton Makrai[2] Balázs Tarján[1] György Szaszák[1]

[1]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics
[2]Research Institute for Linguistics of the Hungarian Academy of Sciences
balint.dobrossy@gmail.com, {tarjanb, szaszak}@tmit.bme.hu, makrai.marton@nytud.mta.hu

Word embeddings and rich morphology?

- Word embeddings
  - represent semantic relations
  - analogical reasoning tasks (Mikolov et al., 2013b; Gladkova and Drozd, 2016)
  - morphosyntax: consistent mapping of grammatical relations
- morphologically rich languages, e.g. Hungarian
  - many word forms
  - less constrained word order ← dependency relations expressed by case endings
- embeddings for rich morphology
  - morphosyntax quite good
  - semantic accuracy of word embedding analogies drops by 49-75% compared to English

What helps?

- vocabulary needs to be increased to ensure a high enough coverage

→ larger training corpus required
- increasing the size of the context window
  - but it may introduce higher context variability
- fastText (Bojanowski et al., 2017) adds character n-grams
- for many languages, this improves both semantic and syntactic accuracy
- no highly agglutinative language tested

Experiments

- analogy set for Hungarian (Makrai, 2015)
  - designed following Mikolov et al. (2013a)
- sub-word unit based embedding strategies
  word embedding trained on the corpus with words divided into segments (as if they were separate words)
- character n-grams
  - baseline Word vectors (trained with fastText)
  - Lemmatization provided by the NLP-pipeline magyarlánc (supervised learning)
  - segments from unsupervised learning (Morfessor): Root or Morf
- different embedding dimensions
- different context window sizes

Corpus, segmentation, and embeddings

- corpus
  - a contemporary dump of Hungarian web pages constructed for this paper, mostly online newspapers in various fields from years 2014–2018
  - over 70 M word tokens
  - also allows for augmentation with character n-grams
- true morphological analysis: magyarlánc (Zsibrita et al., 2013)
  - provides lemmatization in the form of a stem plus a suffix series
  - disambiguation
- unsupervised pseudo-morphemic analysis: Morfessor (Virpioja et al. (2013), morfs)
- text normalization is performed with a Python script
- training the word vector models: fastText (Joulin et al., 2016),

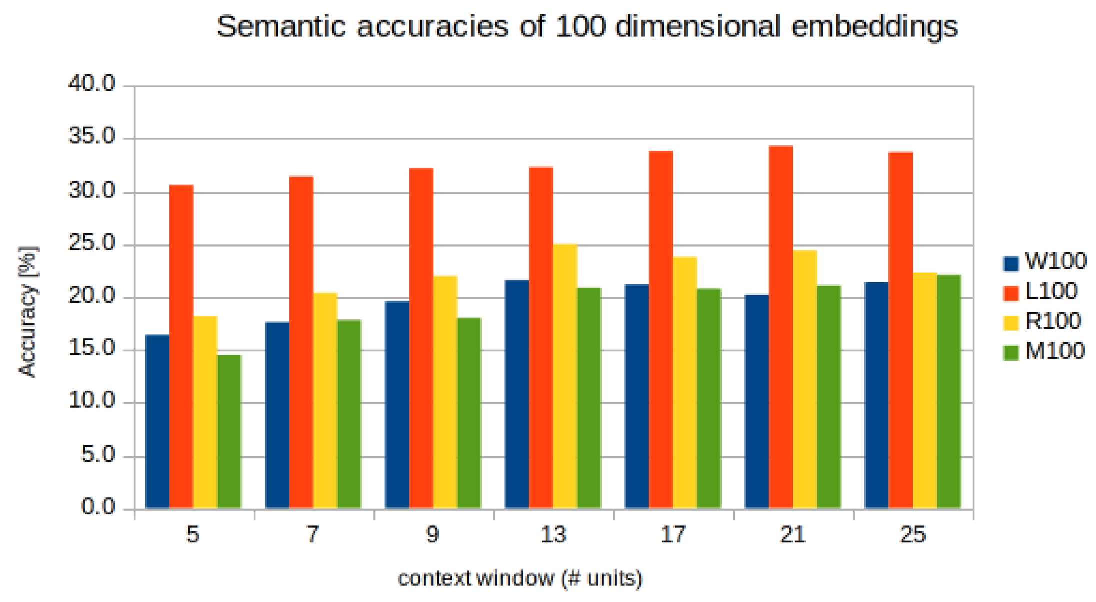
fastText settings

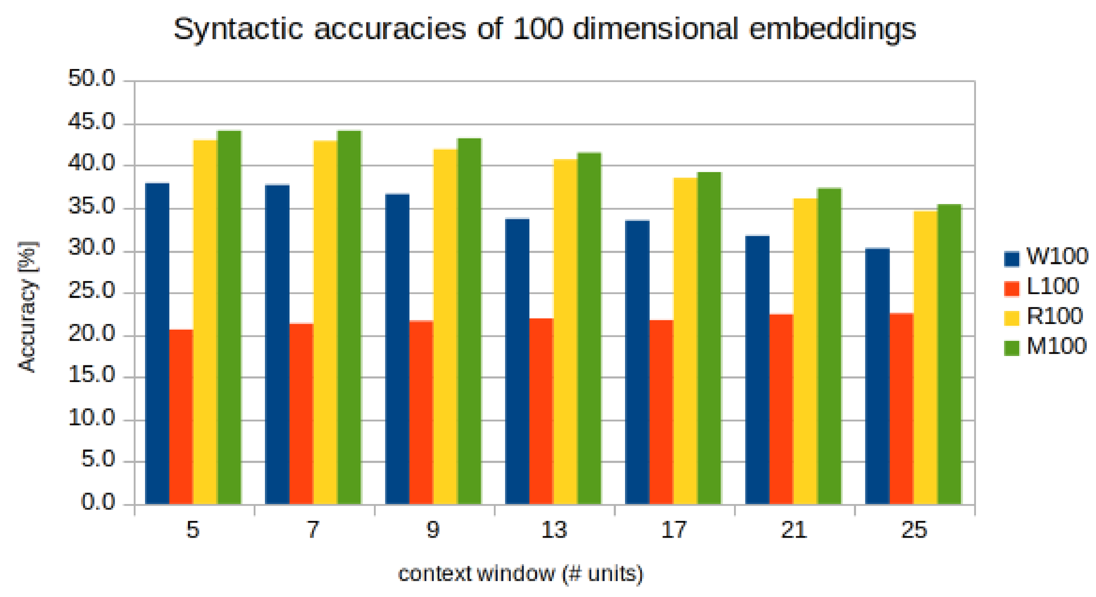
- three main parameters controlled during the experiments:
  - whether we use character n-gram augmentation or not;
  - the size of the context window; and
  - the target dimension of the resulting embedding vectors
- all other parameters at their default value

Parameter	Value range
Frequency cut-off	5
Min length of char ngram	0 or 3
Max length of char ngram	none or 6
Embedding dimension	100-200
Context window	5–25
Learning rate (α)	0.05
α update interval	100
Number of epochs	15
Negative sampling loss	yes
Negative samples	5
Pretraining	none

Extending the context window

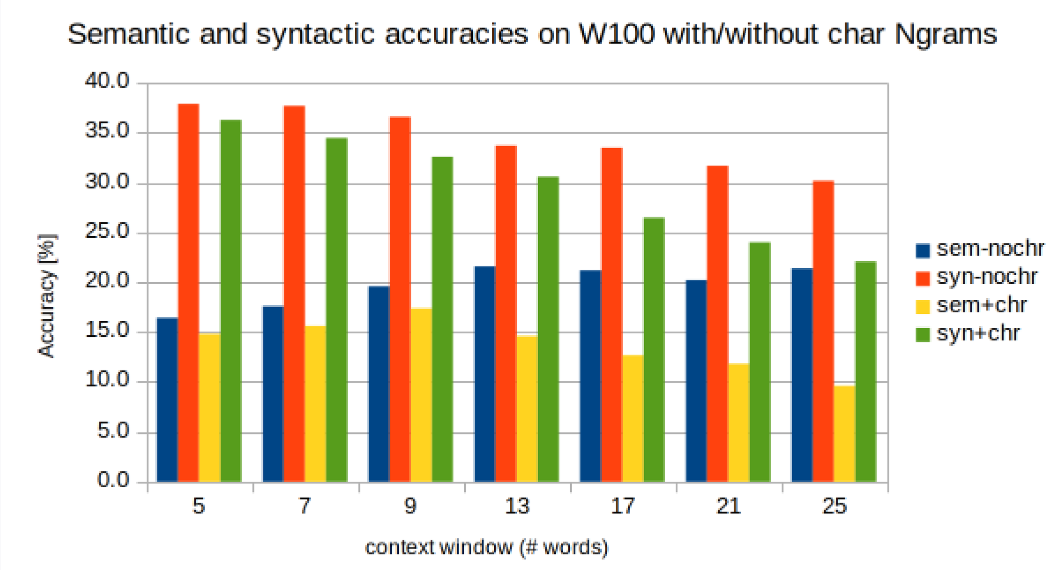
- related work
  - semantic analogical questions benefit from larger windows, syntactic ones do not (Lebret and Collobert, 2015)
  - with SVD models and different window sizes (Gladkova and Drozd, 2016),
    - analogical questions best detected with window size 2–4
    - some questions are equally good at larger windows
    - no one-on-one correspondence between semantics and larger windows





- semantic relations
  - strategies: lemma (L) yields the highest accuracy, 75% higher compared to W
  - long context windows are better (all the four strategies)
- syntactic relations
  - accuracies decrease tendentiously when extending the context window

character n-grams consistently harmful



- both semantic and syntactic accuracy gets lower
- semantic accuracies: no benefit with any of the 4 investigated embedding strategies
- syntax: helpful in some cases (L100, L200 and R200)
- semantics improves with a large window, while morphosyntax does not

Embedding dimension

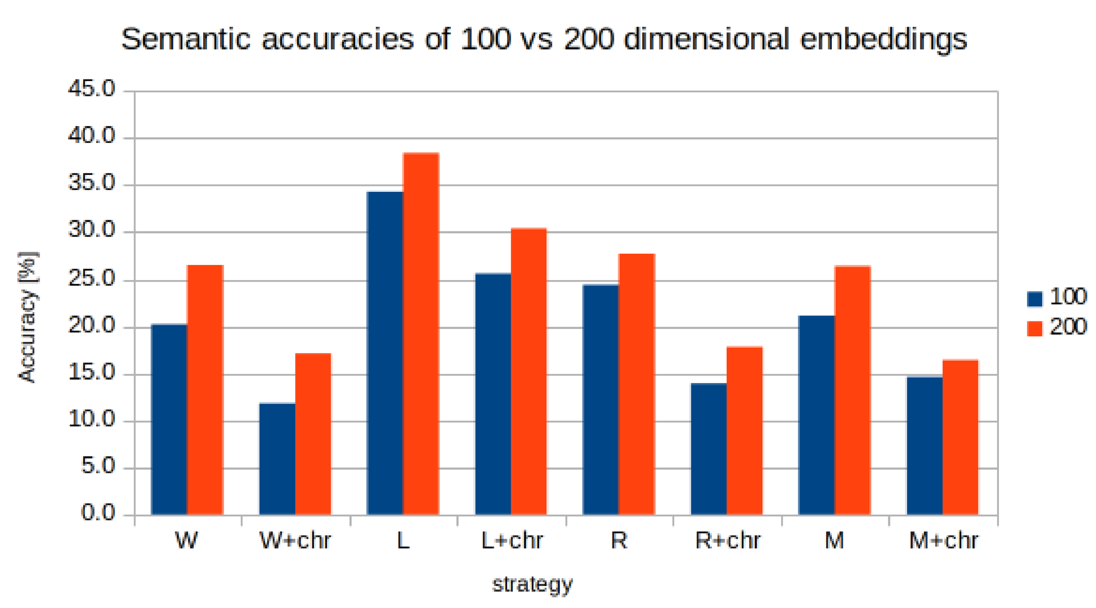


Figure: Context window covers 21 units.

- increasing the embedding dimension helps semantic accuracies: up to 50% relative increase in accuracy
- higher dimensions not considered to avoid making down-stream applications heavy

Individual semantic relations

capital-common-countries	66.0% (101/153)
capital-world	40.3% (2595/6441)
county-center	18.2% (12/66)
currency	6.4% (26/406)
family	16.5% (15/91)
Semantic	38.41% (2749/7157)

Table: Best settings (magyarlánc, window 21, dimension 200, no character n-grams).

Analogical questions

- syntactic and semantic task: Hungarian analogy test (Makrai, 2015)
  - constructed according to (Mikolov et al., 2013a)
- for the semantic accuracy, we use `country-capital` and `currency`
- for the syntactic accuracy we use
  - `gram8-plural-nouns`
  - `gram7-past-tense`
  - `gram3-comparative`
- during testing in analogical questions, query words are also spitted to segments
  - vectors computed as the sum of the segments' vectors
- the semantic part of the Mikolov-style analogical questions focus on named entities
- It is questionable how appropriate it is to use them for the evaluation of the embedding strategies, especially that of encoding lexical semantic relations and not the world knowledge

Related work

- recent study of subword models for morphologically rich languages (Zhu et al., 2019)
  - performance is both language- and task-dependent
  - they miss Hungarian
- Recursive Neural Network (Lazaridou et al., 2013; Luong et al., 2013)
  - morphologically compositional word embeddings, supervised
- analogical questions revisited (Gladkova and Drozd, 2016)
  - different systems shine at different sub-categories of the morphological and semantic tasks
  - derivational morphology is significantly more difficult than inflectional morphology
  - new test set: more difficult
- *byte-pair encoding* (Sennrich et al., 2016)
  - particularly useful for machine translation
- models for many applications augmented with subword in the form of a convolutional neural network or a BiLSTM
- understanding linguistic knowledge encoded in sentence and word embedding modules of
  - neural machine translation (NMT) encoders and decoders
  - deep NLP models (Peters et al., 2018; Smith, 2019)
- individual neurons in deep NLP models Dalvi et al. (2019)
  - linguistic correlation analysis task investigates sensitivity for word-structure (morphology) among other linguistic properties
- Morfessor for automatic speech recognition in rich morphology (Enarvi et al., 2017)
- de-glutinatve method (Borbély et al., 2016; Nemeskey, 2017): inflectional prefixes split into separate tokens for better morphological generalization
- Lévai and Kornai (2019) analyze Hungarian word embedding vectors grouped by the morphological tag
  - Does the coherence of these classes correlate with the specificity or the frequency of the tag?

Future work

- other embedding algorithms
  - besides fastText, the original and the enhanced (Mikolov et al., 2018) word2vec and the GloVe (Řehůřek and Sojka, 2010) implementations of the *continuous bag of words* and the *skip-gram* models
- extend dimensionality up to a few hundred dimensions
- other morphologically rich languages (e.g Finnish, Turkish, or Slavic languages)
  ← translate analogical questions

Acknowledgments

This work was supported by the Hungarian National Research, Development and Innovation Office under contract ID FK-124413: 'Enhancement of deep learning based semantic representations with acoustic-prosodic features for automatic spoken document summarization and retrieval'. Márton Makrai was partially supported by project found 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence and National Research, Development and Innovation Office grant #120145.