Unsupervised Morphological Segmentation Using Neural Word Embeddings

Ahmet Üstün[1] and Burcu Can[2]

[1] Cognitive Science Department, Informatics Institute Middle East Technical University (ODTÜ)
ustun.ahmet@metu.edu.tr

[2] Department of Computer Engineering, Hacettepe University
burcucan@cs.hacettepe.edu.tr

12.10.2016

Contents

What is Morphology?

- Words are made of smaller meaning bearing units which are called *"morphemes"*.

- Morphological segmentation is the process of segmenting words into their morphemes.
    - Stem: <u>advance</u> + ment
    - Suffix: politic + <u>al</u>
    - Prefix: <u>dis</u> + close

What is Morphotactics ?

- Morphotactics involves a set of rules that define *how morphemes can be attached* to each other.

- In agglutinating languages (Turkish, Finnish or Hungarian), concatenation of morphemes plays an important role in morphology.

Related Work on Unsupervised Morphological Segmentation

- Research based on *word-level orthographic patterns*:
  - Goldwater et al. (2006)
  - Creutz and Lagus (2005, 2007)
- Research based on *relation between morphology and syntax*:
  - Can and Manandhar (2010)
- Research based on *relation between morphology and semantics*:
  - Schone and Jurafsky (2001)
  - Narasinham et al. (2015)

Main Intuition

Integrates morphotactics with semantics

Main Intuition

Integrates morphotactics with semantics

↓

Putting semantics at the center of morphology learning task

Main Intuition

Integrates morphotactics with semantics

↓

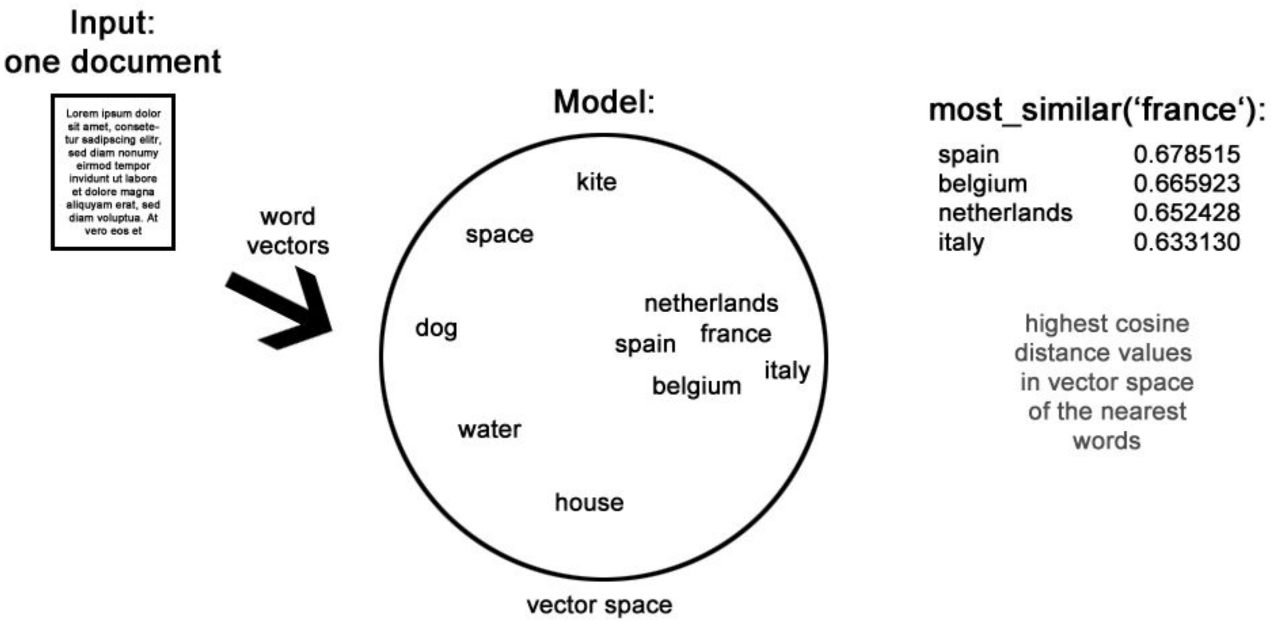
Putting semantics at the center of morphology learning task

↓

Directly using semantic similarity between words to detect segmentation points.

Neural Word Embeddings

- We obtain word embeddings from a raw corpus by using Mikolov et al. (2013)'s *"word2vec"* model in 200 dimensional vector space



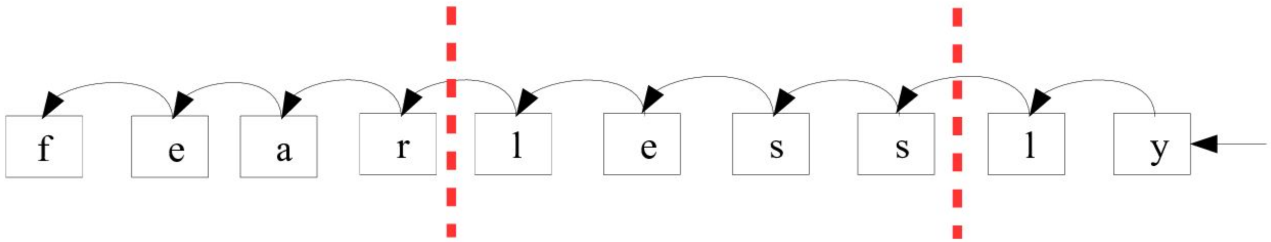
Morphological Segmentation Using Semantic Similarity

- Baseline splitting algorithm is based on the *semantic similarity* between word and its substrings.

- Semantic similarities are obtained by calculating the *cosine distance* between the word embeddings:

$$\cos(v(w_1), v(w_2)) = \frac{v(w_1) \cdot v_i(w_2)}{\|v(w_1)\| \cdot \|v_i(w_2)\|} = \frac{\sum_{i=1}^{n} v_i(w_1) \cdot v_i(w_2)}{\sqrt{\sum_{i=1}^{n} v_i(w_1)^2} \cdot \sqrt{\sum_{i=1}^{n} v_i(w_2)^2}}$$

Morphological Segmentation Using Semantic Similarity



	Word	Remaining substring	Cosine similarity	Segmentation
1	fearlessly	fearlessl	-1	fearlessly
2	fearlessly	fearless	0.34	fearless-ly
3	fearless	fearles	0.14	fearless-ly
4	fearless	fearle	-1	fearless-ly
5	fearless	fear	0.26	fear-less-ly
6	fear	fea	-1	fear-less-ly
7	fear	fe	-1	**fear-less-ly**

Modeling Morphotactics with ML Estimate

- We use *"maximum likelihood estimation (ML)"* to build a bigram language model for morpheme transition.

- ML is modelled according to following formulas:

$$\arg\max_{w=m_0+\cdots+m_N \in W} P(w = m_0 + m_1 + \cdots + m_N) = p(m_0) \prod_{i=1}^{N} p(m_i|m_{i-1}) \quad (1)$$

$$p(m_0) = \frac{n(m_0)}{K} \quad (2)$$

$$p(m_i|m_{i-1}) = \frac{n(< m_i, m_{i-1} >)}{M} \quad (3)$$

Modeling Morphotactics with ML Estimate

- Morphotactics ML model is built on the baseline results that are obtained via semantic splitting algorithm.
- After model training is finished, final segmentation of a word is selected among all possible segmentations by the viterbi algorithm.
- Laplace smoothing with additive number 1 is used to overcome the sparsity problem.

Full Model

Training Word Embeddings with Raw Corpus by Word2Vec

göz	<1 3 4 2 5 ...>
gözle	<5 8 6 1 3 ...>
gözler	<4 1 5 3 9 ...>

.
.

Full Model

Training Word Embeddings with Raw Corpus by Word2Vec → Build Baseline Model by Semantic Splitting Algorithm

göz	<1 3 4 2 5 ...>
gözle	<5 8 6 1 3 ...>
gözler	<4 1 5 3 9 ...>
.	
.	

→

göz-le-r-i-n
(ice, to monitor, his/her ice, your ice)

Full Model

Training Word Embeddings with Raw Corpus by Word2Vec → Build Baseline Model by Semantic Splitting Algorithm → Train Bigram Language Model with ML Estimate for Morphotactics

göz	<1 3 4 2 5 ...>
gözle	<5 8 6 1 3 ...>
gözler	<4 1 5 3 9 ...>
.	
.	

→

göz-le-r-i-n
(ice, to monitor, his/her ice, your ice)

→

göz-ler-in
(your ice)

Data and Parameters

- In order to train word embeddings model:
  - Turkish BOUN corpus: *361M word token, 725K word types*
  - English wiki corpus: *129M word token, 218K word types*

- Baseline semantic splitting algorithm applied on MorphoChallenge (2010) data:
  - Turkish: *617K word token*
  - English: *878K word token*

- For evaluation:
  - Turkish: *1760 word token*
  - English: *1050 word token*

Data and Parameters

- Data composition for whole process is as follows:

	English	Turkish
Word Embeddings	129M	361M
Semantic Parsing and ML Estimation	878K	617K
Development	694	763
Test and Evaluation	1050	1760

Data and Parameters

- In semantic splitting algorithm, we assign cosine similarity threshold as *d = 0.25* to decide the correct split points by performing our models on the development set.

Threshold (d)	Semantic Parsing (%)	Full Model (%)
0.15	40.51	47.51
0.25	37.42	47.82
0.35	30.16	43.58
0.45	25.14	39.95

Results on Turkish

- Comparison of our models with other systems is as follows:

Model	Precision (%)	Recall (%)	F1-measure (%)
Morfessor CatMap	79.38	31.88	45.49
Full Model	50.70	40.07	44.76
Morpho Chain	69.63	31.73	43.60
Aggressive Comp.	55.51	34.36	42.45
Semantic Parsing	61.82	25.42	36.03
Iterative Comp.	68.69	21.44	32.68
Morfessor Baseline	87.35	18.03	29.89
Nicolas	79.02	19.78	31.64
Base Inference	72.81	16.11	26.38

Results on English

- Comparison of our models with other systems is as follows:

Model	Precision (%)	Recall (%)	F1-measure (%)
Morfessor Baseline	66.30	41.28	50.88
Semantic Parsing	64.85	37.75	47.72
Full Model	62.79	35.40	45.28
Morfessor CatMap	64.44	34.34	44.81

Correct and Incorrect Segmentations

- On Turkish:

Correct segmentations	Incorrect segmentations
patlıcan-lar-ı	tiy-at-ro-lar-da
su-lar-da-ki	gaze-t-e-ci-ydi
balkon-lar-da	sipari-ş-ler-i-n-iz
parti-si-ne	gelişti-ril-ir-ken
varis-ler-den	anla-ya-mıyo-r-du-m
entari-li-nin	uygu-lama-sı-nda-n
üye-ler-i-dir-ler	veri-tabanları-yla

- On English:

Correct Segmentations	Incorrect segmentations
vouch-safe-d	cen-tr-alize-d
dictator-ial	ni-hil-ist-ic
help-less-ness	su-f-fix-es
rational-ist	ba-ti-ste
express-way	sh-o-gun
flow-chart	el-e-v-ation-s
drum-head-s	im-pe-rsonator-s

Correct and Incorrect Segmentations

- On Turkish:

Correct segmentations	Incorrect segmentations
patlıcan-lar-ı	tiy-at-ro-lar-da
su-lar-da-ki	gaze-t-e-ci-ydi
balkon-lar-da	sipari-ş-ler-i-n-iz
parti-si-ne	gelişti-ril-ir-ken
varis-ler-den	anla-ya-mıyo-r-du-m
entari-li-nin	uygu-lama-sı-nda-n
üye-ler-i-dir-ler	veri-tabanları-yla

Main problem is oversegmentation !

Conclusion and Future Work

We presented:

- Probabilistic model that integrates morphotactics with word embeddings to use semantics in morphological segmentation task
- Especially in agglutinating languages our model performs challenging results.

Future Work:

- Joint model that learns morphology, syntax and dependency structure with the help of sematics.

References

- Can, B., Manandhar, S.: Clustering morphological paradigms using syntactic categories. In: Multilingual Information Access Evaluation I. Text Retrieval Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers. pp. 641-648. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
- Clark, A.: Inducing syntactic categories by context distribution clustering. In: Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7. pp. 91-94. ConLL'00, Association for Computational Linguistics, Stroudsburg, PA, USA (2000)
- Creutz, M., Lagus, K.: Inducing the morphological lexicon of a natural language from unannotated text. In: Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005. pp. 106{113 (2005)
- Creutz, M., Lagus, K.: Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report A81 (2005)
- Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions Speech Language Processing 4, 3:1-3:34 (February 2007)
- Goldwater, S., Griffths, T.L., Johnson, M.: Interpolating between types and tokens by estimating power-law generators. In: Advances in Neural Information Processing Systems 18. p. 18 (2006)
- Hankamer, J.: Finite state morphology and left to right phonology. In: Proceedings of the Fifth West Coast Conference on Formal Linguistics (January 1986)
- Kurimo, M., Lagus, K., Virpioja, S., Turunen, V.T.: Morpho challenge 2010. http://research.ics.tkk.fi/events/morphochallenge2010/ (June 2011), online; accessed 4-July-2016

References

- Lee, Y.K., Haghighi, A., Barzilay, R.: Modeling syntactic context improves morphological segmentation. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 1-9. CoNLL '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
- Lignos, C.: Learning from unseen data. In: Kurimo, M., Virpioja, S., Turunen, V., Lagus, K. (eds.) Proceedings of the Morpho Challenge 2010 Workshop. pp. 35-38. Aalto University, Espoo, Finland (2010)
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
- Narasimhan, K., Barzilay, R., Jaakkola, T.S.: An unsupervised method for uncovering morphological chains. Transactions of the Association for Computational Linguistics (TACL) 3, 157167 (2015)
- Nicolas, L., Farre, J., Molinero, M.A.: Unsupervised learning of concatenative morphology based on frequency-related form occurrence. In: Kurimo, M., Virpioja, S., Turunen, V., Lagus, K. (eds.) Proceedings of the Morpho Challenge 2010 Workshop. pp. 39-43. Aalto University, Espoo, Finland (2010)
- Schone, P., Jurafsky, D.: Knowledge-free induction of in ectional morphologies. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. pp. 1-9. NAACL'01, Association for Computational Linguistics, Stroudsburg, PA, USA (2001)
- Soricut, R., Och, F.: Unsupervised morphology induction using word embeddings. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. pp. 1627-1637 (2015)
- Sproat, R.W.: Morphology and computation. MIT press (1992)
- Team, D.D.: Deeplearning4j: Open-source distributed deep learning for the jvm, apache software foundation license 2.0. http://deeplearning4j.org/ (May 2016)

Questions ?