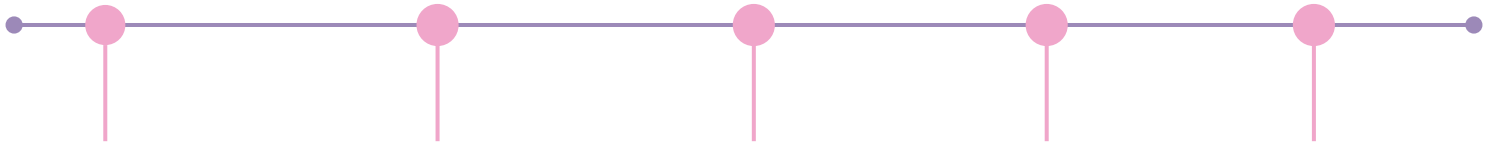
Multilingual substitutes in WSI

Князьковой Виктории Игоревны

https://github.com/vknyazkova/Multilinguial_Substitutes_WSI

Table of contents



Word Sense Induction and Lexical Substitution approach

Word occurrences:

cell.n.74: *It's in **cell** phones and laptop computers.*
cell.n.76: *Whether it is **cell** phones, digital cameras or flat TV screens, Japan continues to dominate many sectors of technology…*
cell.n.75: *From the moment you pick up a lab report stating that the doctor has diagnosed cancerous **cells**, life starts to get tough*
cell.n.78: *The free radicals are released when oxygen carrying red blood **cells** called haemoglobin die*

Clustering:

Word Sense Induction and Lexical Substitution approach

Word substitutes:

cell.n.74: *It's in **[phone cellphone phone-like]** phones and laptop computers.*
cell.n.76: *Whether it is **[cellphone phone phone-like]** phones, digital cameras or flat TV screens, Japan continues to dominate many sectors of technology…*
cell.n.75: *From the moment you pick up a lab report stating that the doctor has diagnosed cancerous **[tumor carcinoma cancer-derived]**, life starts to get tough*
cell.n.78: *The free radicals are released when oxygen carrying red blood **[blood-cell red-blood lymphocyte]** called haemoglobin die*

Clustering:

Word Sense Induction and Multilingual Lexical Substitution approach

Word substitutes:

cell.n.74: *It's in **[phone телефон Telefon]** phones and laptop computers.*

cell.n.76: *Whether it is **[cellphone мобильный Handy]** phones, digital cameras or flat TV screens, Japan continues to dominate many sectors of technology…*

cell.n.75: *From the moment you pick up a lab report stating that the doctor has diagnosed cancerous **[tumor рак Knochenmark]**, life starts to get tough*

cell.n.78: *The free radicals are released when oxygen carrying red blood **[blood-cell лимфоциты Lungenzellen]** called haemoglobin die*

Clustering:

Experimental setup: Dataset

<cell.n.16>The team used a battery of the newly developed `` gene probes , '' snippets of genetic material that can track a gene 's presence in a cell .
<TargetSentence>By analyzing cells extracted from eye tumors , they found defects in the second copy of chromosome 13 in the exact area as in the first copy of the chromosome . </TargetSentence>The finding riveted medicine .
</cell.n.16>

Semeval 2010 Task 14

- 100 words
- 50 nouns
- 50 verbs
- Up to 3 context sentences

Experimental setup: Languages + model

Languages

en, ru, fr, es, de

Masked language model

bert-base-multilingual-uncased

Substitutes Generation

Instance

*Boy I'm **rooting** for George Mason*

→ **Context translation**

Парень, я болею за Джорджа Мейсона

→ **Word alignment**

*Boy I'm **rooting** for George Mason*

*Парень, я **болею** за Джорджа Мейсона*

→ **Context-based probability**

*Парень, я **[MASK]** за Джорджа Мейсона*

Target

*bert.embedding[**root**]*

→ **Proximity to the target**

$$P(s|T) \propto \frac{exp(< emb_s, emb_T >)}{\tau}$$

+embs

Probability distribution

$$P(s|C,T) \propto P(s|C)P(s|T)$$

https://aclanthology.org/2020.coling-main.107/

class BertProbEstimator

Parameters: *temperature*

- **context_conditioned_probs**
- **target_conditioned_probs**

class MultilingualSubstituteGenerator

Parameters:

languages
top_k
target_injection
source_lang_inj
words_only_subst
lemmatize

source_lang_inj

In +embs strategy which target inject - original or translated

words_only_subst

При ранжировании подстановок закрывать маской сабтокены и пунктуацию или нет

Clustering

01 Vectorization

Vectorize substitutes for every instance of target

02 Clustering algorithms

Either Agglomerative clustering or K-means

03 Number of clusters

Different strategies for selecting number of clusters

Clustering: vectorization

tf-idf

Не учитывает ранг подстановки

access.n.1: ‘provide allow able’
access.n.2: ‘able get allow’

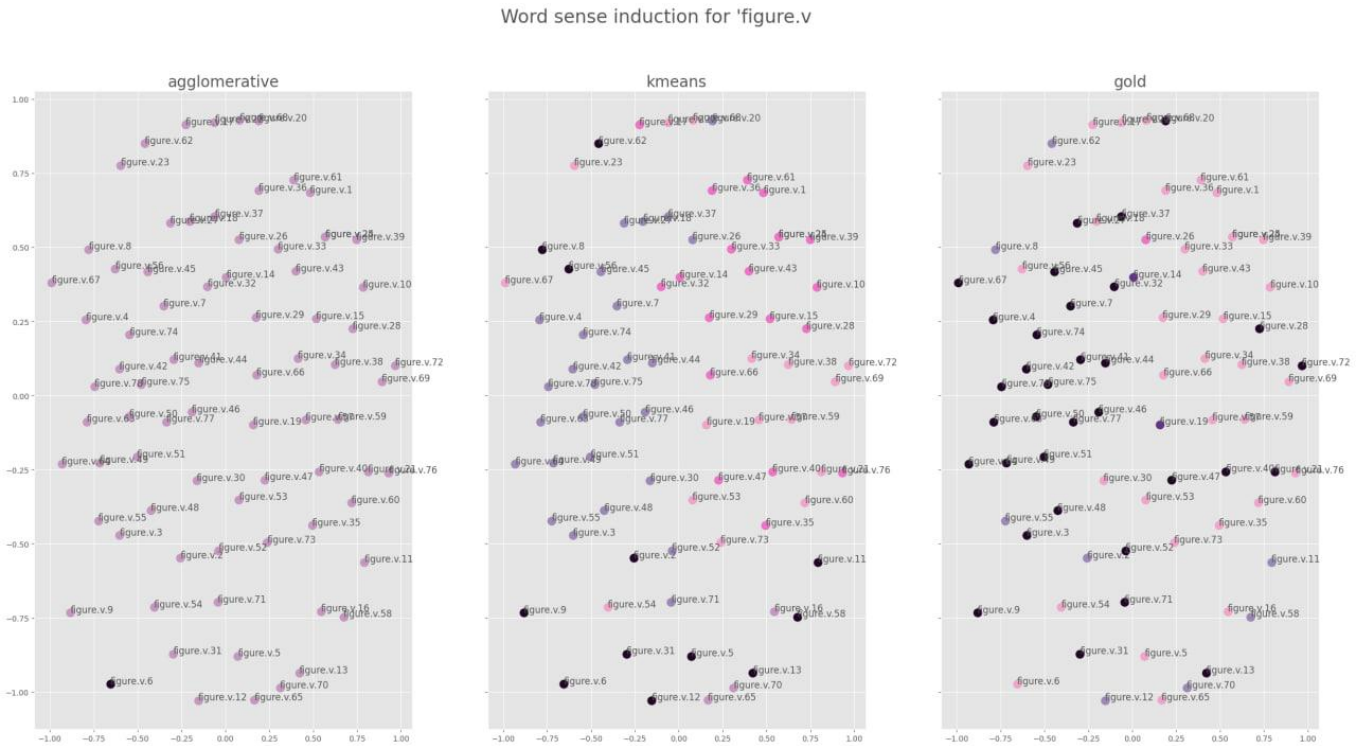
tf-idf-weighted

Вес подстановок зависит от ее ранга в списке замен

access.n.1: ‘provide provide provide allow allow able’
access.n.2: ‘able able able get get allow’

\+ *both methods **without** idf*

Clustering: algorithms



Agglomerative

faster
better overall results

but: *unbalanced clusters*

K-means

much slower
worse mean metrics

but: *more plausible clusters*

Clustering: number of clusters

fix

Predetermined number of clusters for all target words

Maximizing silhouette score

Selecting from some range number of clusters, that maximizes silhouette score

Predetermined range

Default: range(2, 10)

From 2 to number of contexts

For every target its own range

Experiments: target injection, clusterization

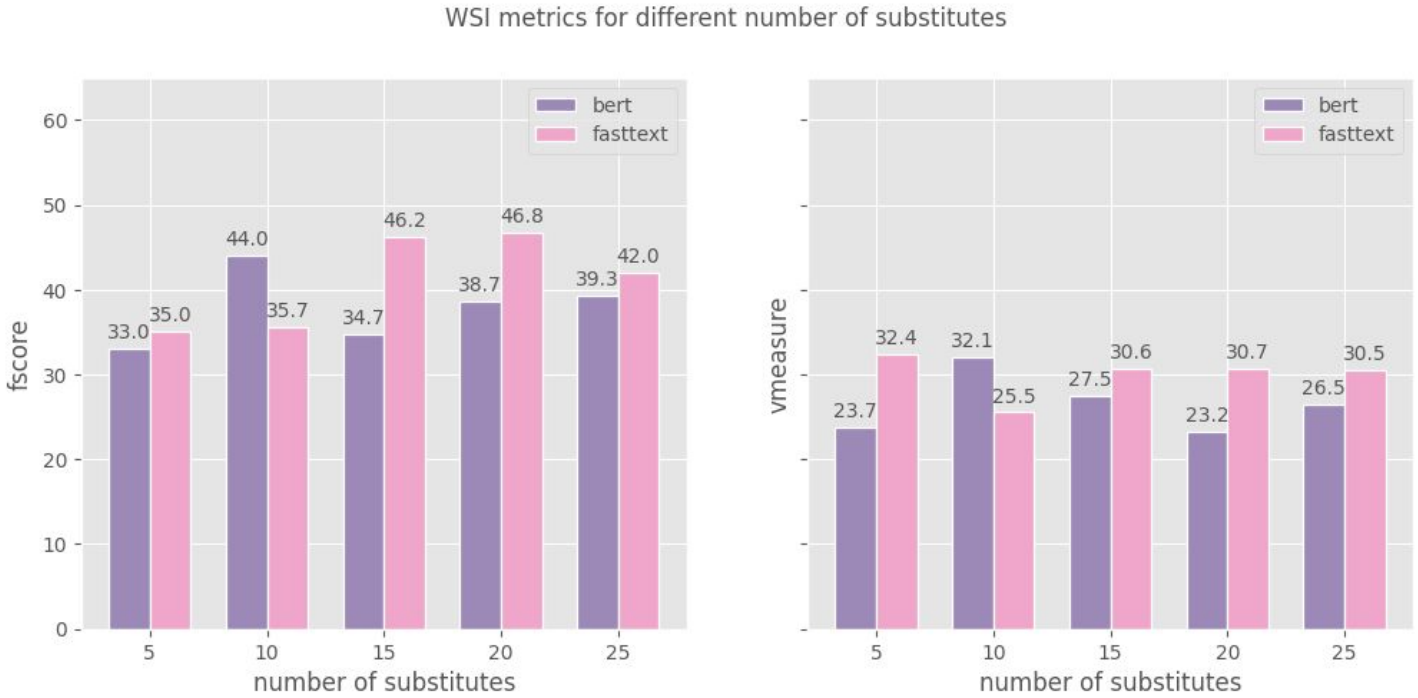
Best mean metric: *agglomerative with* ***source language*** *+embs*

target language	clusterizer	fscore	precision	recall	vmeasure	homogenity	completeness	(fs * vm) ** 0.5
original	kmeans	42.47	36.73	63.86	28.75	24.69	43.49	32.661893
original	agglomerative	52.32	54.58	59.65	32.89	36.34	38.01	39.568463
translation	kmeans	32.29	26.11	58.58	27.61	23.56	42.88	29.186564
translation	agglomerative	48.31	45.58	61.83	29.47	26.83	39.95	35.558408

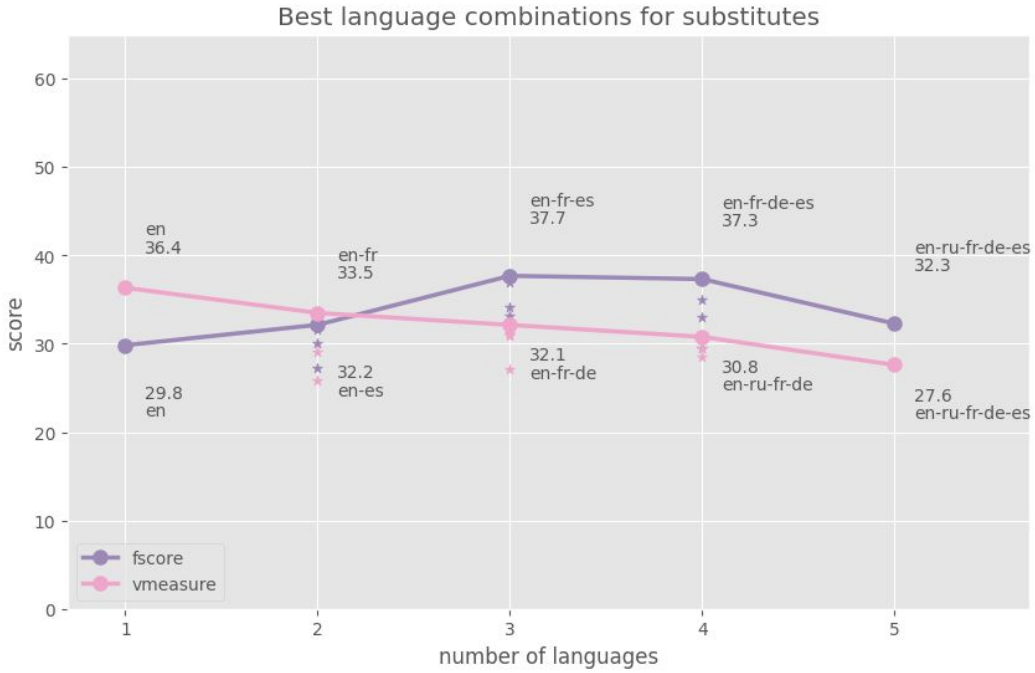
Experiments: number of substitutes

Fasttext best: *15 substitutes*
Bert best: *10 substitutes*

But: *overall bert worse :(*



Results: Different combinations



*Incorporation of additional languages **can improve** english-only result (a little)*

*Some **language combinations** give better results than others*

Conclusions

Multilingual substitutes **can** possibly **improve** WSI results

Но это улучшение **незначительно**, причем оказалось, что это не особо зависит от архитектуры модели

Возможно, стоит переосмыслить весь пайплайн (• ᴖ • ｡)

Appendix: semeval-2010 results

System	VM (%) (All)	VM (%) (Nouns)	VM (%) (Verbs)	#Cl
Hermit	16.2	16.7	15.6	10.78
UoY	15.7	20.6	8.5	11.54
KSU KDD	15.7	18	12.4	17.5
Duluth-WSI	9	11.4	5.7	4.15
Duluth-WSI-SVD	9	11.4	5.7	4.15
Duluth-R-110	8.6	8.6	8.5	9.71
Duluth-WSI-Co	7.9	9.2	6	2.49
KCDC-PCGD	7.8	7.3	8.4	2.9
KCDC-PC	7.5	7.7	7.3	2.92
KCDC-PC-2	7.1	7.7	6.1	2.93
Duluth-Mix-Narrow-Gap	6.9	8	5.1	2.42
KCDC-GD-2	6.9	6.1	8	2.82
KCDC-GD	6.9	5.9	8.5	2.78
Duluth-Mix-Narrow-PK2	6.8	7.8	5.5	2.68
Duluth-MIX-PK2	5.6	5.8	5.2	2.66
Duluth-R-15	5.3	5.4	5.1	4.97
Duluth-WSI-Co-Gap	4.8	5.6	3.6	1.6
Random	4.4	4.2	4.6	4
Duluth-R-13	3.6	3.5	3.7	3
Duluth-WSI-Gap	3.1	4.2	1.5	1.4
Duluth-Mix-Gap	3	2.9	3	1.61
Duluth-Mix-Uni-PK2	2.4	0.8	4.7	2.04
Duluth-R-12	2.3	2.2	2.5	2
KCDC-PT	1.9	1	3.1	1.5
Duluth-Mix-Uni-Gap	1.4	0.2	3	1.39
KCDC-GDC	7	6.2	7.8	2.83
MFS	0	0	0	1
Duluth-WSI-SVD-Gap	0	0	0.1	1.02

System	FS (%) (All)	FS (%) (Nouns)	FS (%) (Verbs)	#Cl
MFS	63.5	57.0	72.7	1
Duluth-WSI-SVD-Gap	63.3	57.0	72.4	1.02
KCDC-PT	61.8	56.4	69.7	1.5
KCDC-GD	59.2	51.6	70.0	2.78
Duluth-Mix-Gap	59.1	54.5	65.8	1.61
Duluth-Mix-Uni-Gap	58.7	57.0	61.2	1.39
KCDC-GD-2	58.2	50.4	69.3	2.82
KCDC-GDC	57.3	48.5	70.0	2.83
Duluth-Mix-Uni-PK2	56.6	57.1	55.9	2.04
KCDC-PC	55.5	50.4	62.9	2.92
KCDC-PC-2	54.7	49.7	61.7	2.93
Duluth-WSI-Gap	53.7	53.4	53.9	1.4
KCDC-PCGD	53.3	44.8	65.6	2.9
Duluth-WSI-Co-Gap	52.6	53.3	51.5	1.6
Duluth-MIX-PK2	50.4	51.7	48.3	2.66
UoY	49.8	38.2	66.6	11.54
Duluth-Mix-Narrow-Gap	49.7	47.4	51.3	2.42
Duluth-WSI-Co	49.5	50.2	48.2	2.49
Duluth-Mix-Narrow-PK2	47.8	37.1	48.2	2.68
Duluth-R-12	47.8	44.3	52.6	2
Duluth-WSI-SVD	41.1	37.1	46.7	4.15
Duluth-WSI	41.1	37.1	46.7	4.15
Duluth-R-13	38.4	36.2	41.5	3
KSU KDD	36.9	24.6	54.7	17.5
Random	31.9	30.4	34.1	4
Duluth-R-15	27.6	26.7	28.9	4.97
Hermit	26.7	24.4	30.1	10.78
Duluth-R-110	16.1	15.8	16.4	9.71