EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Word Sense Disambiguation using Word Embeddings Information

Monday Seminar at UFAL

Ebrahim Ansari

25/02/2019

Outline:

- Word Sense Disambiguation (WSD)
- Word representations
- Unsupervised Approach
- Supervised Approach
- Conclusion

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Word Sense Disambiguation (WSD)

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Human language is inherently ambiguous!

English	Persian
Bank (n) : 1. Financial Institution 2. Riverside …	آن یکی شیر است اندر بادیه آن دگر شیر است اندر بادیه آن یکی شیر است کآدم می خورد و آن دگر شیر است کآدم میخورد

Term	Meaning	Spelling	Pronunciation
Synophone	Different	Different	Similar but not identical
Synonym	Same	Different	Different
Polyseme	Different but related	Same	Same or different
Homophone	Different	Same or different	Same
Homonym	Different	Same	Same
Homograph	Different	Same	Same or different
Heteronym	Different	Same	Different
Heterograph	Different	Different	Same
Capitonym	Different when capitalized	Same except for capitalization	Same or different

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

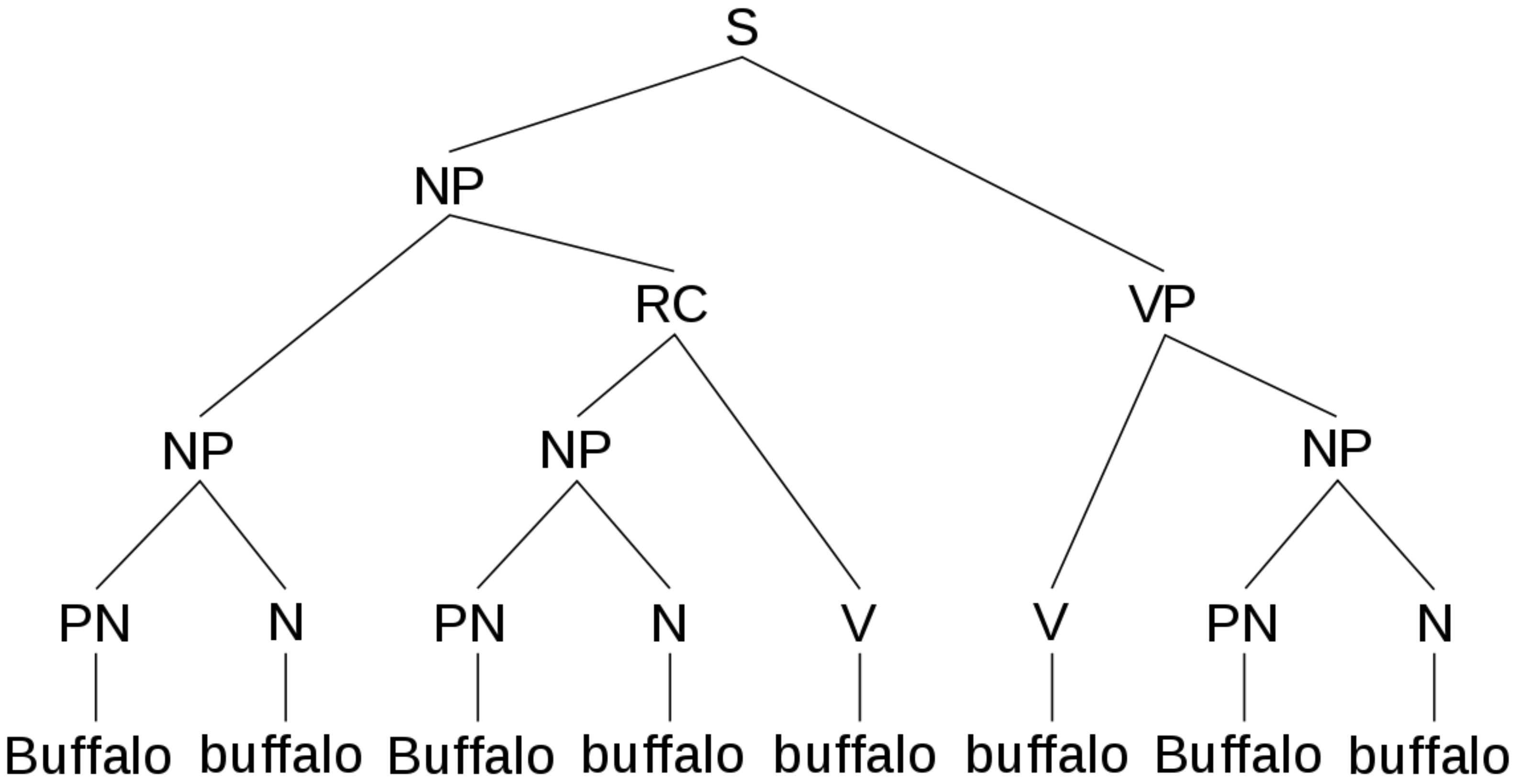
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Word Sense: One of the meanings a word may have depending on the context:

- The man cashed a check at the bank.
- He sat on the bank of the river and watched the currents.

WSD: The process of automatically finding the correct sense of the polysemous words in a given text.

An example: **Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.**



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

WSD applications:

- Machine Translation
- Information Retrieval
- Word processing
- Information extraction and text mining
- Content and sentiment analysis

WSD approaches:

- Supervised
- Unsupervised
- Knowledge based approaches
- Semi-supervised
- Hybrid approaches

Most Frequent Sense (MFS): A baseline

- Among senses of a given word one sense is occurred more than others.
- MFS Identifies the most often used meaning and uses this meaning by default.
- The MFS baseline is often hard to beat for any WSD system and it is considered as the strongest baseline in WSD.
- For example Consider word “شیر” in Persian:

  Translations= {“milk”, “lion”, “faucet”, …}

  MFS (“شیر”) is the Milk translation in English

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Lesk Algorithm (unsupervised)

- Identify senses of words in context using definition overlap.
- Consider two words W_1 and W_2

(1) for each sense *i* of W_1
(2)   for each sense *j* of W_2
(3)   compute *Overlap(i,j)*, the number of words in common
            between the definitions of sense *i* and sense *j*
(4) find *i* and *j* for which *Overlap(i,j)* is maximized
(5) assign sense *i* to W_1 and sense *j* to W_2

Agirre, Eneko, and Philip Edmonds, eds. Word sense disambiguation:
Algorithms and applications. Vol. 33. Springer Science & Business Media, 2007.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Lesk algorithm (example): Consider two words Pine and Cone

- **Pine:**
  1. seven kinds of evergreen tree with needle-shaped leaves
  2. pine
  3. waste away through sorrow or illness
  4. pine for something, pine to do something
- **Cone:**
  1. solid body which narrows to a point
  2. something of this shape, whether solid or hollow
  3. fruit of certain evergreen trees (fir, pine)

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Lesk algorithm (example): Consider two words Pine and Cone

- **Pine:**
  1. **seven kinds of evergreen tree with needle-shaped leaves**
  2. pine
  3. waste away through sorrow or illness
  4. pine for something, pine to do something
- **Cone:**
  1. solid body which narrows to a point
  2. something of this shape, whether solid or hollow
  3. **fruit of certain evergreen trees (fir, pine)**

Overlap(pine1, cone3) = {“evergreen”, ”tree” , “pine”}

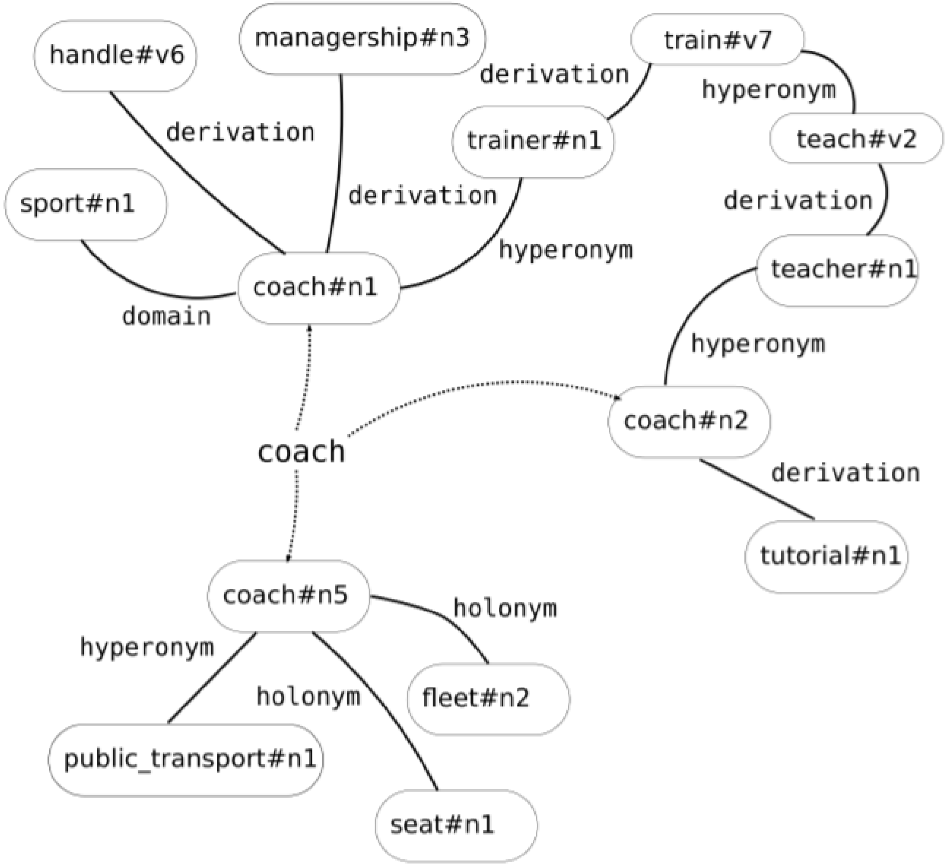
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Graph-based methods (Agirre et al., 2014)

A WSD algorithm based on random walks over large Lexical Knowledge Bases (LKB)

Best results when they used WordNet and eXtended WordNet



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Graph-based methods (Agirre et al., 2014), cont.

- Uses random walk (page-rank)
- Uses WordNet to create graph

$$\mathbf{P} = cM\mathbf{P} + (1-c)\mathbf{v}$$

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Word Representations

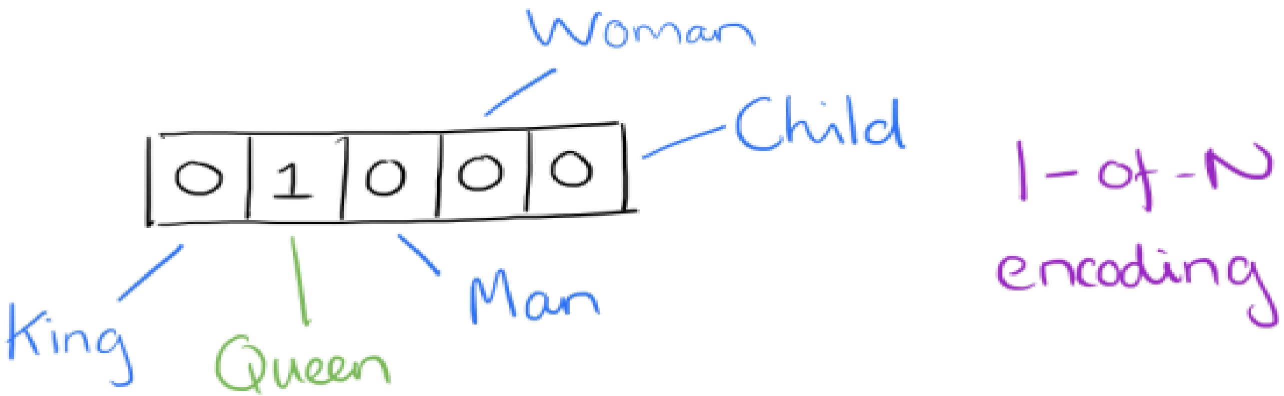
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Word representations – one-hot vectors

Each word in vocabulary is represented with one bit in a huge vector.

- Ex: Hello is [000000100000000] in a vocabulary of size 15.
- No contexts information



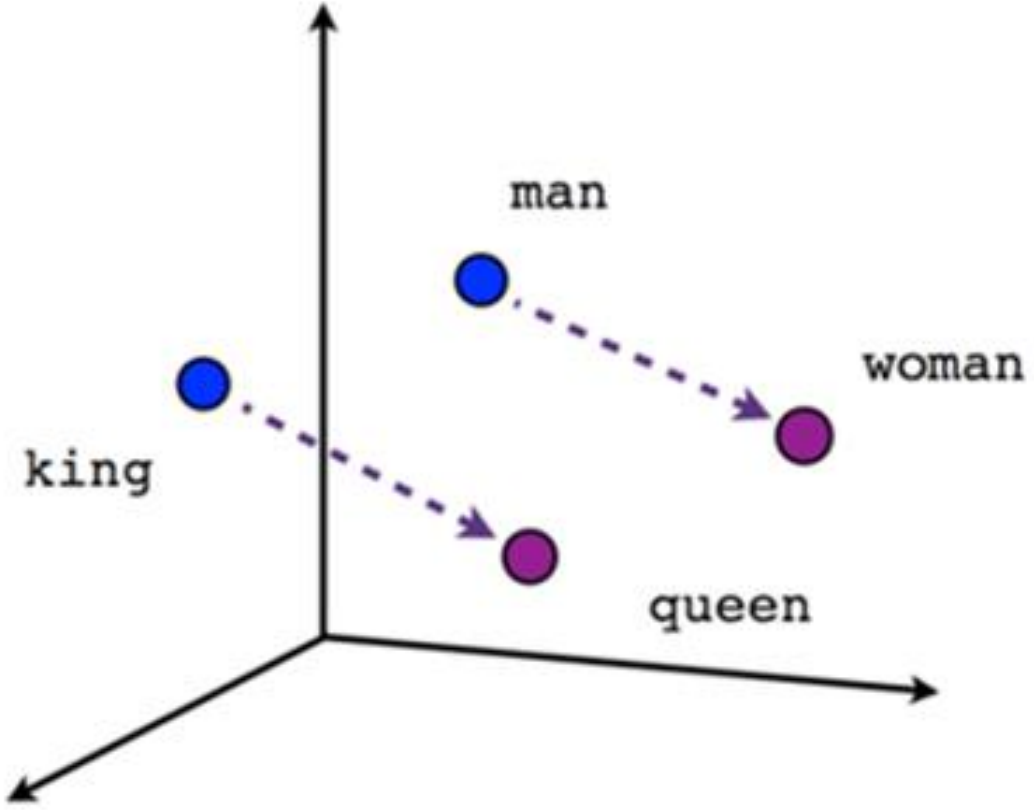
https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Word representations – word embeddings

Each word is represented as a point in a space with fixed number of dimensions

- Ex: Hello can be like [0.4, -0.11, 0.55, 1, …..]

vector[Queen] = vector[King] - vector[Man] + vector[Woman]



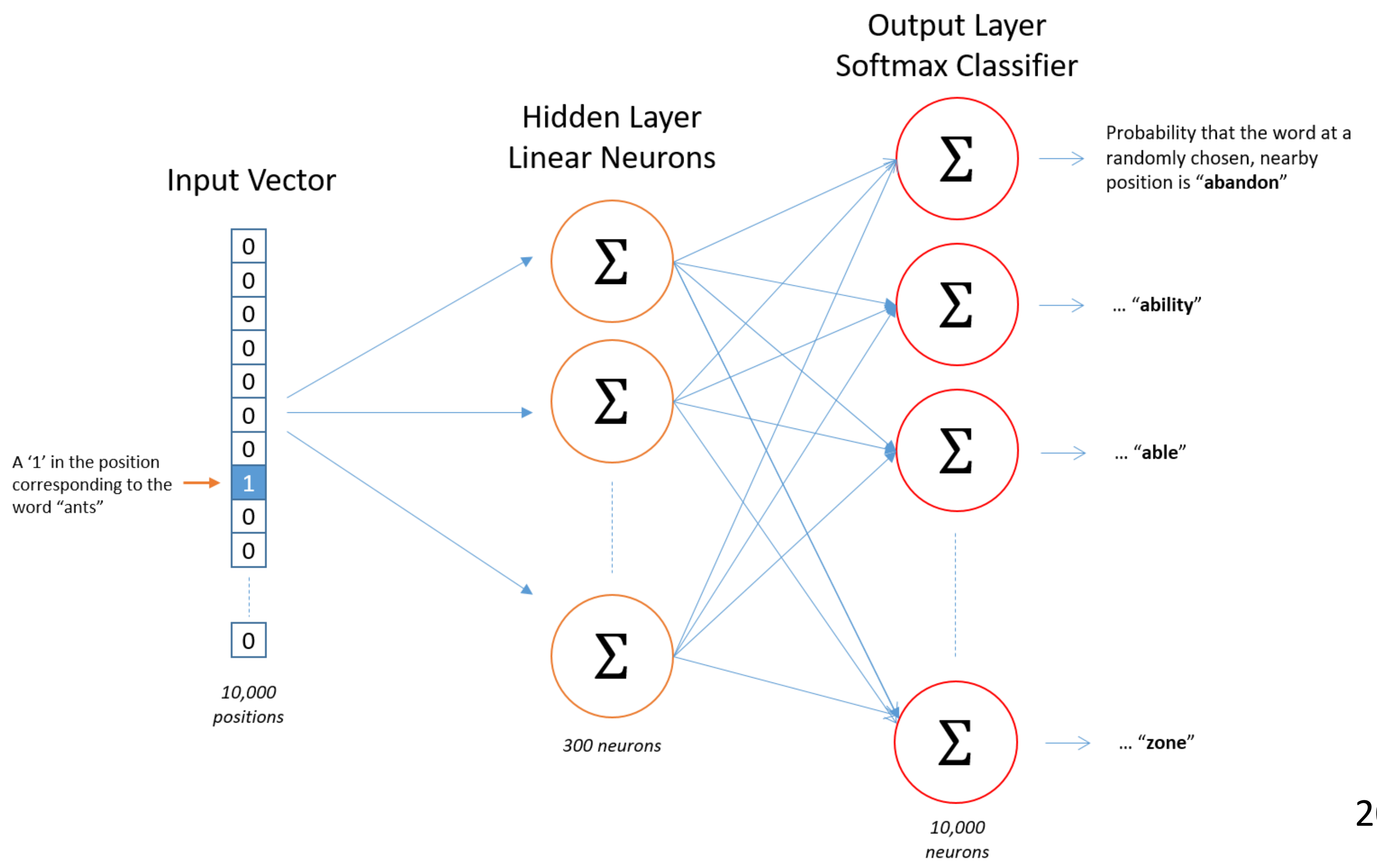
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

word

embeddings

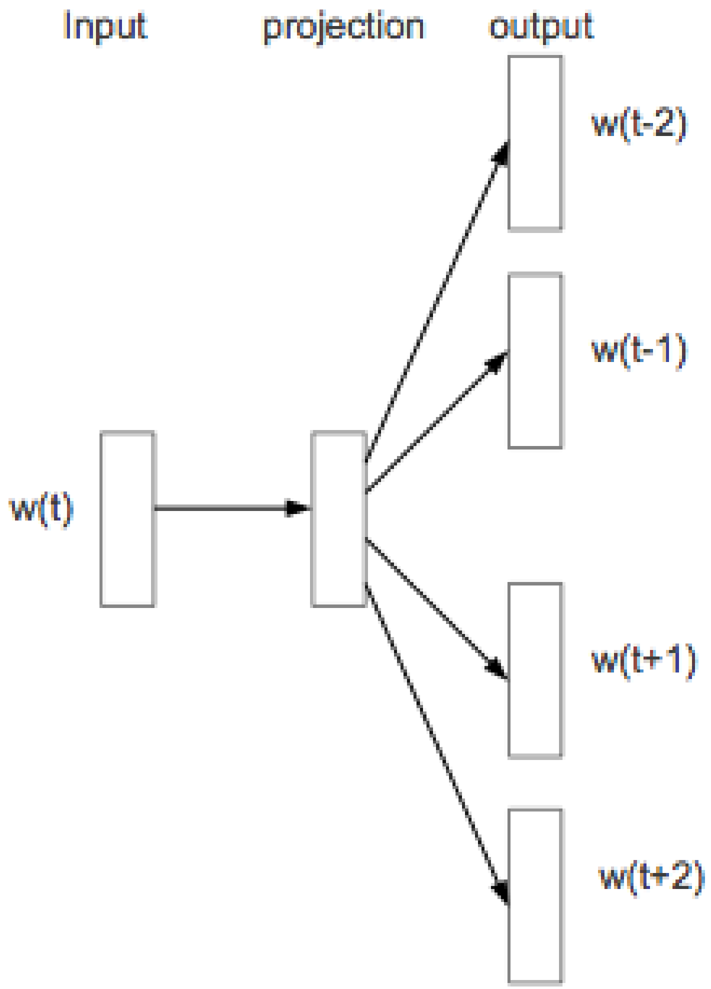
word2vec



word2vec: Skip-Gram

Predict surrounding words

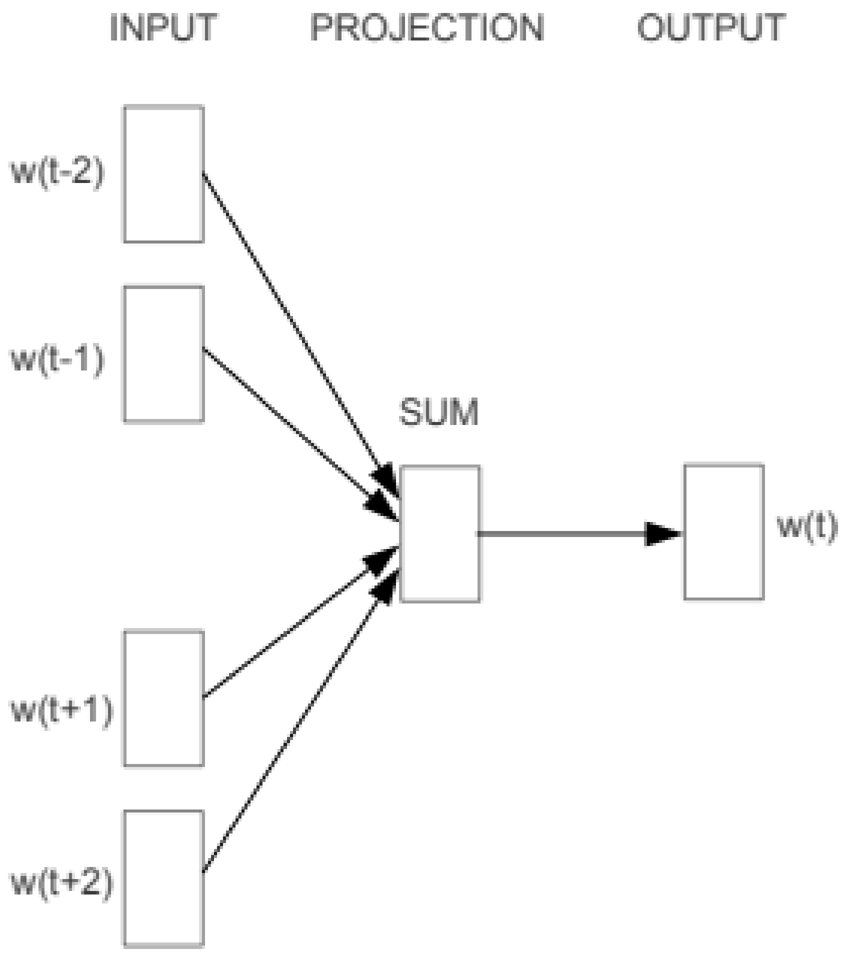
using given word



word2vec: CBOW

Predict current word

given the context



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Unsupervised Approach

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Unsupervised Word Sense Disambiguation using Word Embeddings

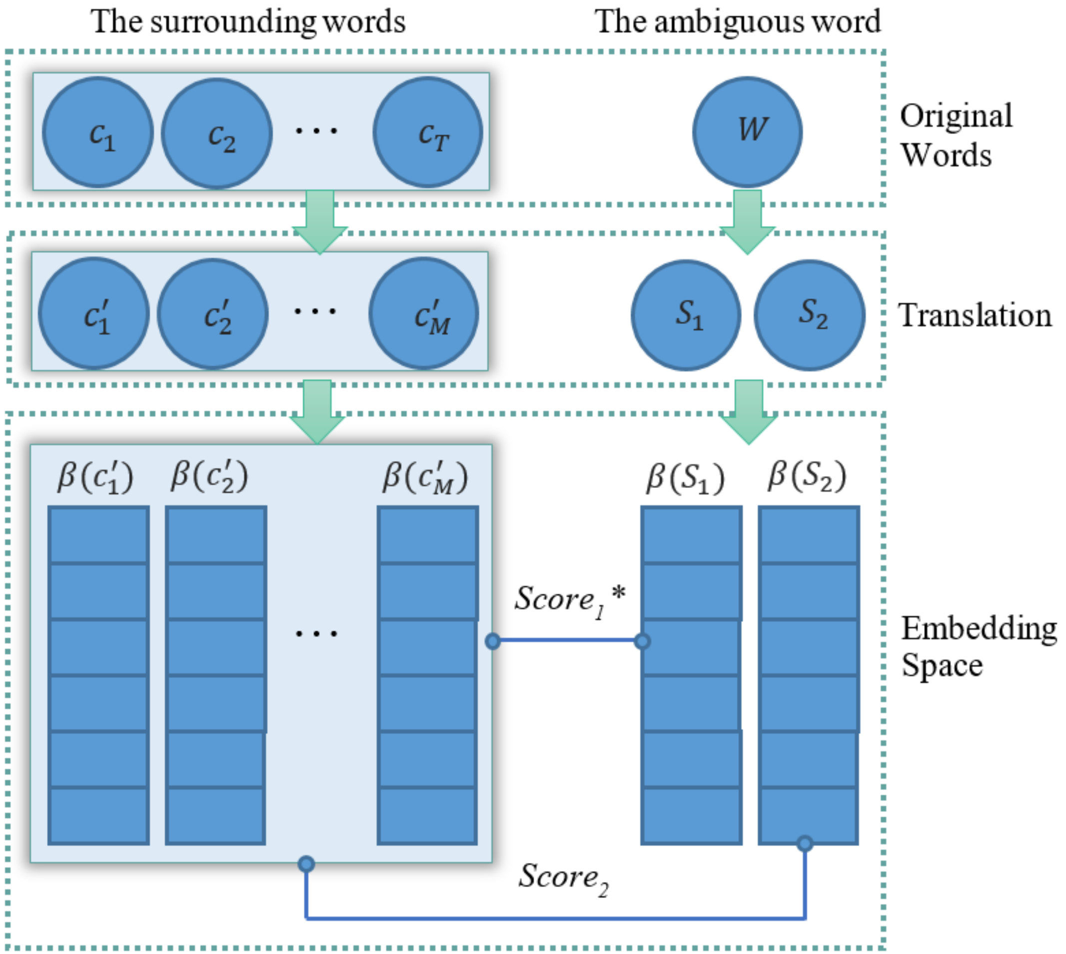
To Disambiguate words from the first language (i.e. Persian) by deploying the trained word embeddings model of the second language (i.e. English) using only a bilingual dictionary

Each translation of the polysemous word is compared against word embeddings of translated surrounding words

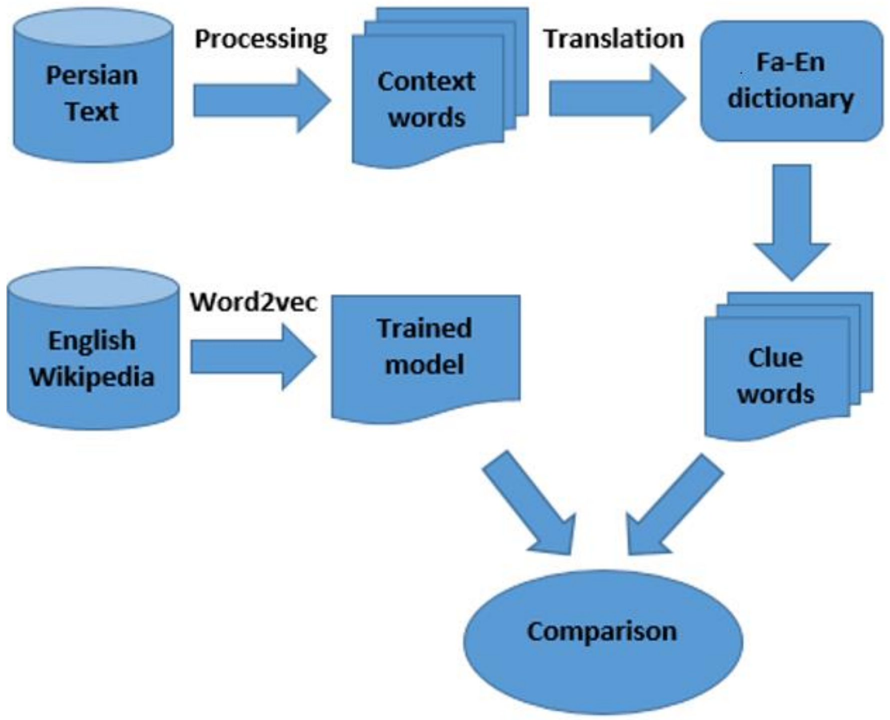
Most similar word to vectors of translated surrounding words is selected as the correct translation

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS



*$Score_i$: the score obtained from comparison between S_i and translated surrounding words



Definitions

C = { c_1, c_2, … , c_T}, context words of w

S = {s_1, s_2, … , s_N}, possible senses of word w

- Using *D* as the bilingual dictionary:

C'={$t_1^1, \ldots, t_1^{N_{c_1}}, t_2^1, \ldots, t_2^{N_{c_2}}, \ldots, t_T^1, \ldots, t_T^{N_{c_T}}$ },

t_j^i represents the *i-th* candidate translation of c_j and N_{c_j} is the possible number of translations of word c_j .

Definitions (cont.)

- For simplicity assume that M=$\sum_{k=1}^{T} N_{c_k}$. Hence, another representation for C' is

$$C' = \{c'_1, c'_2, \dots , c'_M\}$$

- using β (word to vector function)

$$\beta(S) = \{\beta(s_1), \beta(s_2), \dots, \beta(s_N)\}$$

$$\beta(C') = \{\beta(c'_1), \beta(c'_2), \dots , \beta(c'_M)\}$$

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Sum-Vec Strategy (SVS) – first strategy

- Sum vector of vectors within set $\beta(C')$ will be computed (named *R*)
- F_i= f($\beta(s_i)$, R) represents similarity between *i-th* candidate translation and R. Thus the set F= $\{F_1, F_2, \ \dots\ , F_N\ \}$ is provided

s^* =$argmax_{s\in S}$*F*.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Each-Vec Strategy (EVS) – second strategy

- Each vector within β(S) is compared against each vector within $\beta(C')$
- F_i= $\{f_{i1}, f_{i2}, \dots . f_{iM}\}$, where f_{ij}= $f(\beta(s_i), \beta(c_j'))$
- G= $\{G_1, G_2, \dots , G_N\}$ where $G_i = \frac{1}{M} \sum_{j=1}^{M} f_{ij}$ *(average value for* F_i*)*

$s^* = argmax_{s \in S} G$.

Data

- Data for this study were collected from Persian Wikipedia articles containing ambiguous words. Despite the difficulty of creating new test dataset, the producing of data was done manually.

- The dictionary used in this study is a word by word bilingual Persian-English dictionary including about 85K entries of Persian words

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

We use python implementation of *word2vec* embedded in Gensim.

In this experiment eight configurations are selected which are seen here

Configuration	Number of Dimensions	Window Size	Min Count
1	200	5	5
2	200	5	10
3	200	10	5
4	200	10	10
5	400	5	5
6	400	5	10
7	400	10	5
8	400	10	10

List of Configurations

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Best results:

Senses	شیر [*Šir*]		سبک [*Sabok/Sabk*]		جو [*Jo/Jav*]		جرم [*Jorm/Jerm*]	
	Milk	Lion	Style	Light	Atmosphere	Barley	Mass	Crime
# of senses	134	66	138	62	134	66	160	40
# of corrects	126	42	117	46	128	61	158	32
	168		163		189		190	
accuracy	84%		81.5%		94.5%		95%	

Table 4: Results of Each-Vec Strategy for cosine similarity in configuration 1

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

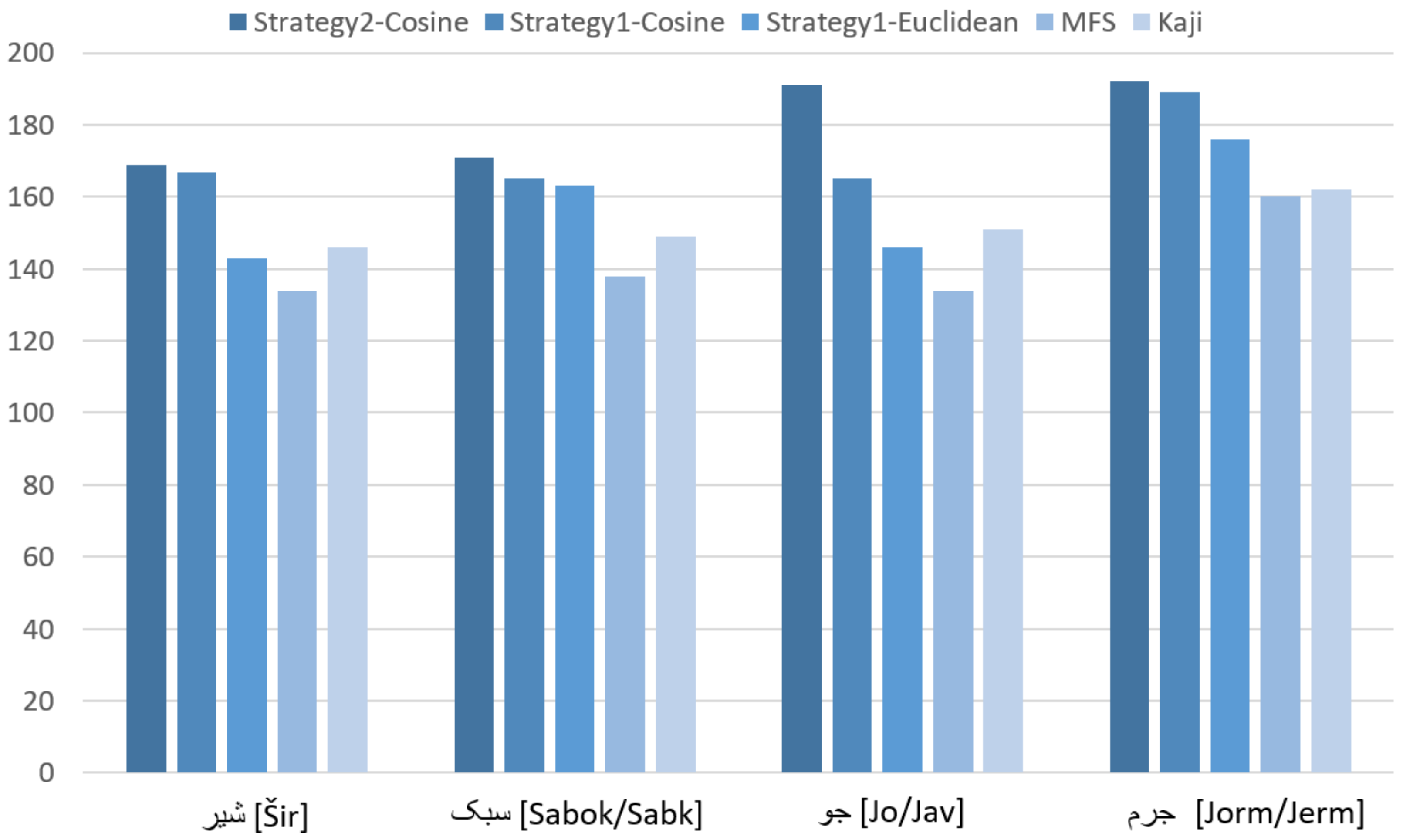
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Comparison:

	شیر [*Šir*]	سبک [*Sabok/Sabk*]	جو [*Jo/Jav*]	جرم [*Jorm/Jerm*]	**Overall Accuracy**
MFS	67%	69%	67%	80%	70.75%
Strategy1-Cosine	82.5%	**81.5%**	88.5%	94.5%	86.75%
Strategy1-Eucledean	77.5%	81%	75.5%	87.5%	80.40%
Strategy2-Cosine	**84%**	**81.5%**	**94.5%**	**95%**	**88.75**

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS



https://iasbs.ac.ir/~ansari/nlp/wsdw2vec.html

A new unsupervised word sense disambiguation for Persain words using comparable corpora

Comaparable corpus

- Click here to download 2016 English wikipedia articles
- Click here to download 2016 Persian wikipedia articles

Dictionary

The dictionary we use is a bilingual word by word Persian-English. It contains 83505 Persian entries with their transaltions.

- You can download Persian-English dictionary here

Test data and Goldtext

For each ambiguous word 200 text samples(paragraphs or simple sentences) are extracted from Persian articles of Wikipedia 2016. Then the ambiguous words were tagged with their sense manaully.

You can download test data for 4 words:

- شیر (Šir)
- سبک (Sabok_Sabk)
- جرم (Jorm_Jerm)
- جو (Jo_Jav)

Also there are some related data for disambiguating Persain words including extracted sentences with annotated ambiguous words from Hamshahri corpus provided by E. Ansari and H.Mousavi.

After preproccessing Hamshahri corpus, sentences containing 8 ambiguous Persian words are extracted then stopwords are removed and all senetnces are saved in xml files according their intended ambigous words

Extracted desired sentences(stopwords are removed):

words	number of sentences	download
اشکال (Aškâl_Eškâl)	7504	download
جو (Jo_Jav)	10021	download

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Supervised Approach

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Supervised Word Sense Disambiguation Using New Features Based on Word Embeddings

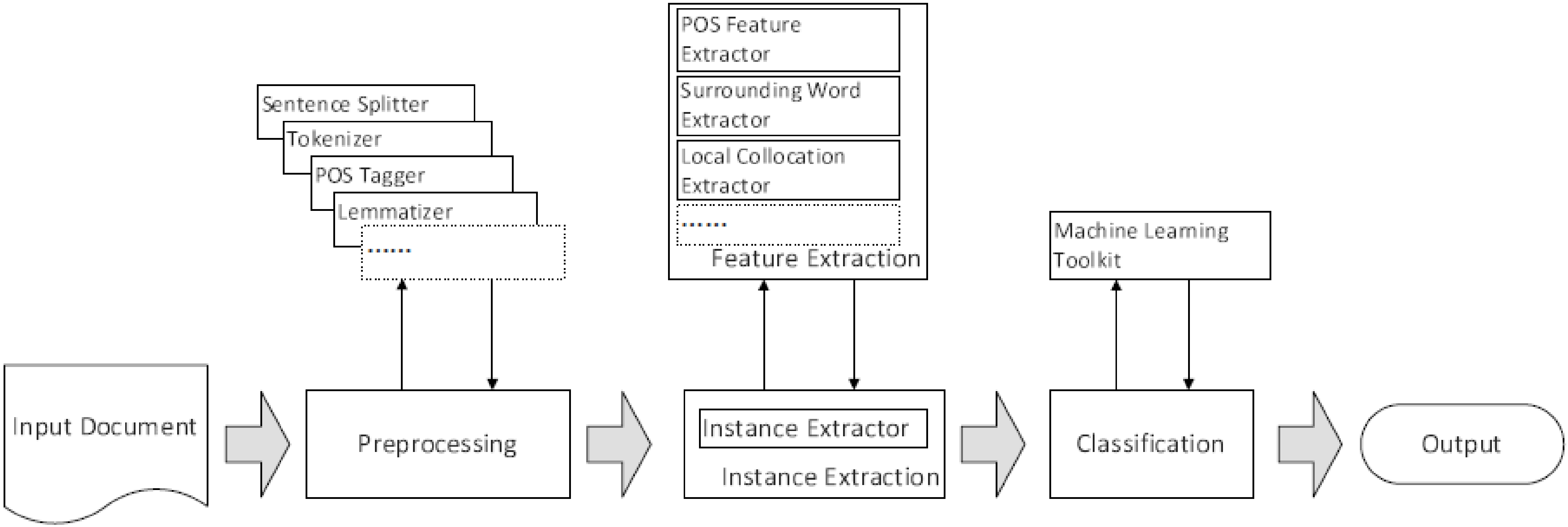
Four improvements to existing state-of-the-art WSD methods

- A new model for assigning vector coefficients
- We applied a PCA dimensionality reduction process
- A new weighting scheme is suggested to tackle the problem of unbalanced data
- A novel voting idea is presented to combine word embedding features extracted from different independent corpora

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

IMS (Zhong and Ng, 2010)



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Iacobacci et al., 2016

- Iacobacci et al. introduced a new method for using word embeddings as features to a WSD system.

We modified this work, proposing four novel ideas which will be discussed in more details in the next section.

Iacobacci et al., 2016 (cont.)

Concatenation $$e_i = \begin{cases} w_{i \bmod D, I-W+\lfloor \frac{i}{D} \rfloor} & if_{\lfloor \frac{i}{D} \rfloor} \langle W \\ w_{i \bmod D, I-W+1} + \lfloor \frac{i}{D} \rfloor & otherwise \end{cases}$$

Average $$e_i = \sum_{\substack{j=I-W \\ j \neq I}}^{I+W} \frac{w_{ij}}{2W}$$

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Iacobacci et al., 2016 (cont.)

Fractional Decay

$$e_i = \sum_{\substack{j=I-W \\ j \neq I}}^{I+W} w_{ij} \frac{W - |\mathrm{I} - j|}{W}$$

Exponential Decay

$$e_i = \sum_{\substack{j=I-W \\ j \neq I}}^{I+W} w_{ij} (1 - \alpha)^{|\mathrm{I} - j| - 1}$$

$$\alpha = 1 - 0.1^{(W-1)^{-1}}$$

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Idea 1 of 4 – Part 1

Using New Coeffs in Exponential Decay Strategy (distance)

$$e_i = \sum_{\substack{j=I-W \\ j \neq I}}^{I+W} dist_{ij} \cdot w_{ij} (1-\alpha)^{|I-j|-1}$$

Idea 1 of 4 – Part 2

Using New Coeffs in Exponential Decay Strategy (count)

$$e_i = \sum_{\substack{j=I-W \\ j \neq I}}^{I+W} count_j \cdot w_{ij} (1-\alpha)^{|I-j|-1}$$

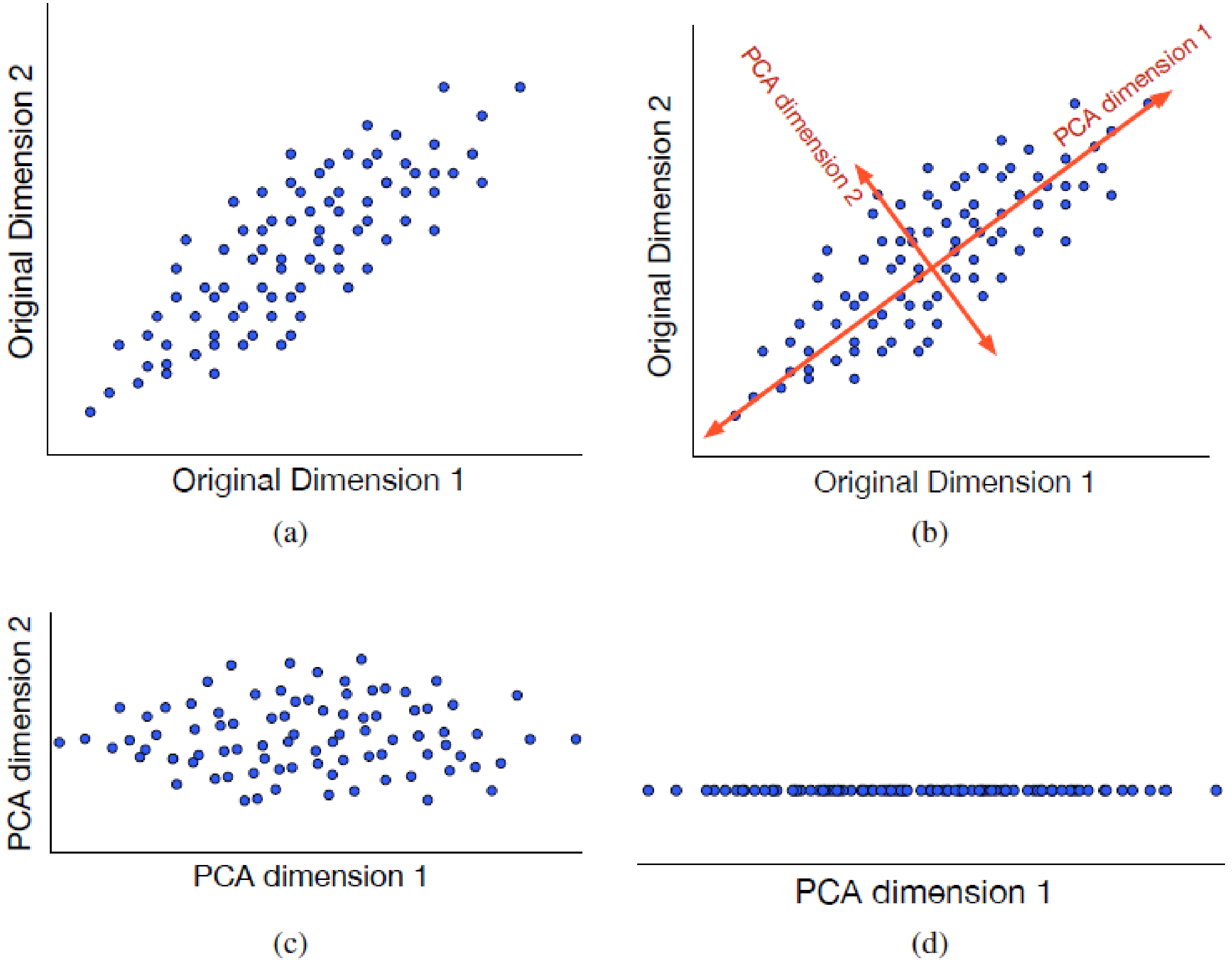
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Idea 2 of 4 – Using PCA

Inspired by the work of Raunak (2017)

Even dimension changing
Leads us better results



Idea 3 of 4: *New Weighting Scheme*

Data is imbalance

Word		Cool (a)	Party (n)
Number of each sense in training set	Sense 1	53	148
	Sense 2	25	15
	Sense 3	3	16
	Sense 4	8	39
	Sense 5	0	17
	Sense 6	1	-
	Sense 7	18	-

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Idea 3 of 4: *New Weighting Scheme (cont.)*

A possible hyper-plane can be represented by:

$$W' \cdot \Phi(x + b) = 0.$$

Where W, is the weight vector normal to the hyperplane and Φ(*x*) is the mapping function that transforms data points to a higher dimensional space.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Idea 3 of 4: *New Weighting Scheme (cont.)*

The maximum margin hyper-plane can be found by solving the following optimization problem

$$min(\frac{1}{2}W \cdot W + C^{+} \sum_{i|y_j=+1}^{l} \zeta_i + C^{-} + \sum_{i|y_j=-1}^{l} \zeta_i)$$

$$s.t. \quad y_i(W \cdot \Phi(x_i) + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0, \quad i = 1, ...l$$

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Idea 3 of 4: *New Weighting Scheme (cont.)*

Akbani et al. argued that by setting $C-/C+$ equal to the minority to majority class ratio, an optimal solution is obtained.

In a multi-label classification task using SVM, the C parameter of each class is computed as follows:

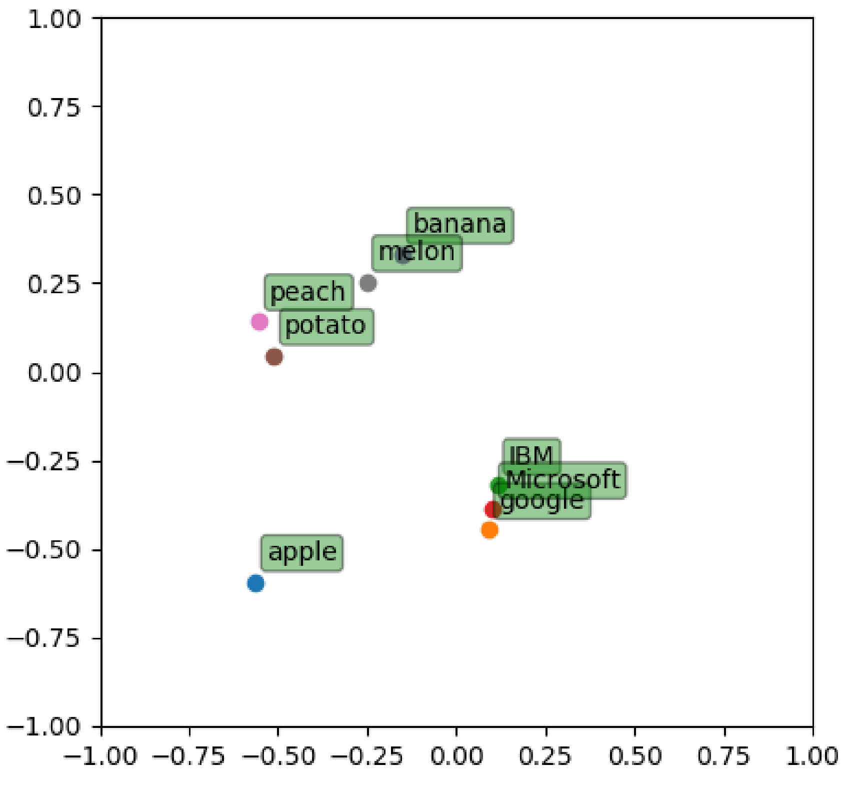
$$C_i = \max(S) \,/\, \mathrm{count}(i)$$

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

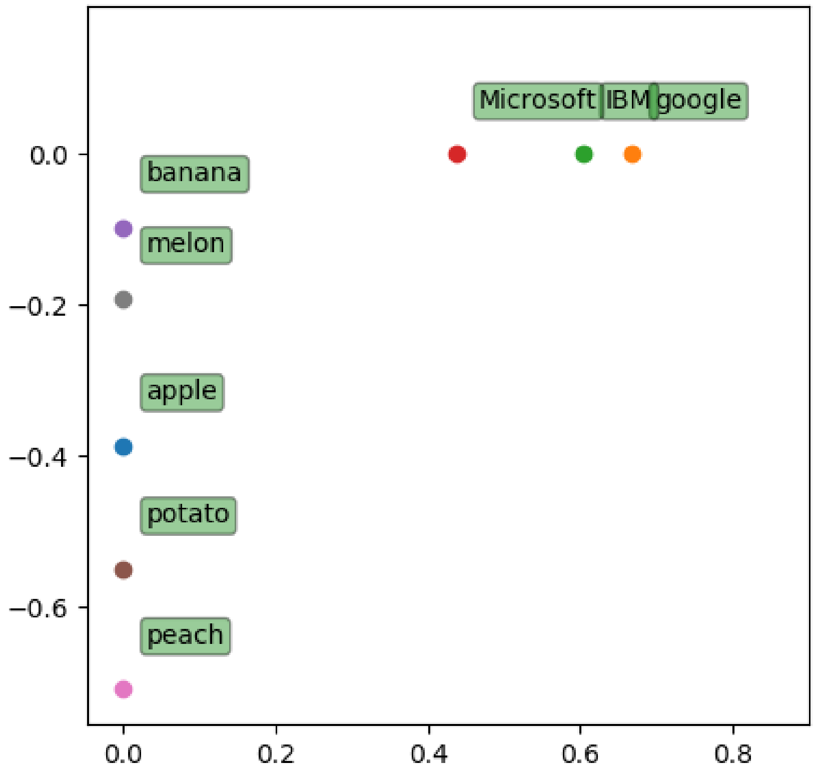
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

- Idea 4 of 4 – *Voting as a Word Embeddings Aggregation Method*

Extracted from Wikipedia



Extracted from Google News

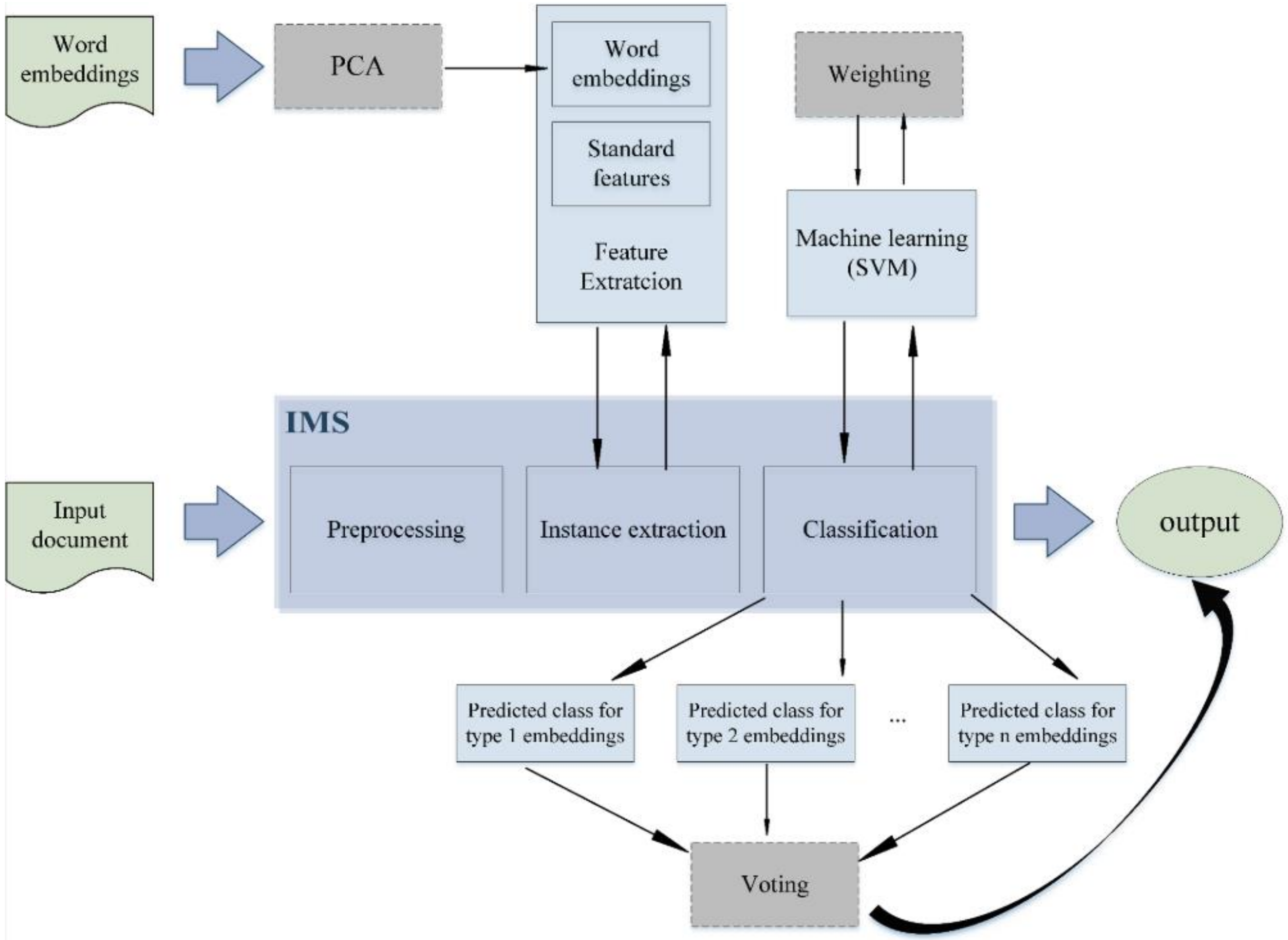


- Idea 4 of 4 – *Voting (cont.)*

$$Sense_v = \arg\max_s f(w, s)$$

$$f(w, s) = \sum_{i=1}^{n} (s_i | s_i \textit{ is the probability of sense s of word w for embedding type i.})$$

where *n* is the total number of embedding types


Word embeddings
PCA
Word embeddings
Standard features
Feature Extratcion
Weighting
Machine learning (SVM)
IMS
Input document
Preprocessing
Instance extraction
Classification
output
Predicted class for type 1 embeddings
Predicted class for type 2 embeddings
...
Predicted class for type n embeddings
Voting

Experiments:

Word Embedding Properties

Word Embeddings	Dimensions	Tokens
Wikipedia 2014	400	1604163
Google news	300	3000000
Fasttext	300	2519370

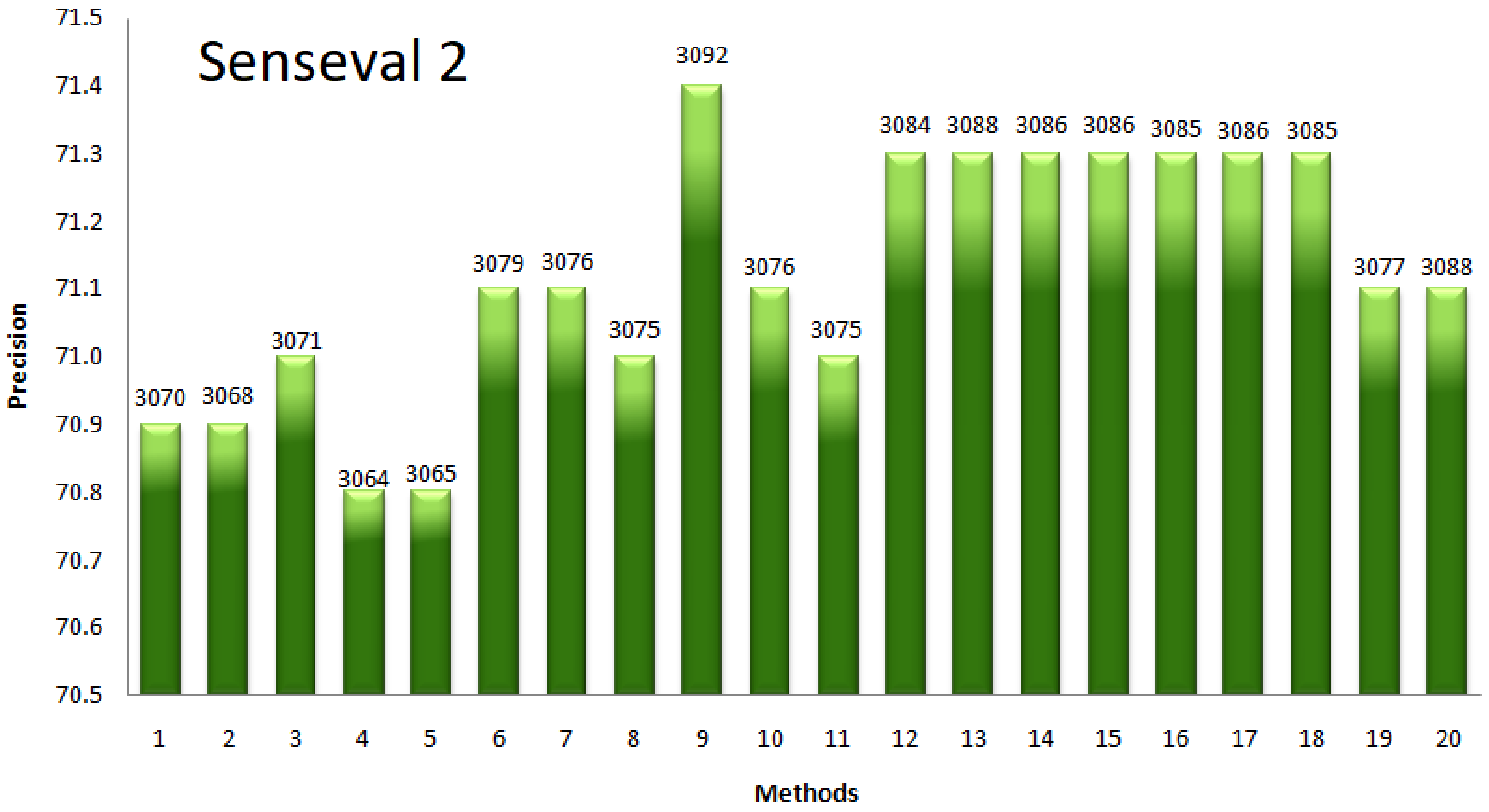
Dataset Description

	Senseval 2	Senseval 3
Word Types	73	57
Training Samples	8611	8022
Test Samples	4328	3944

		Senseval2		Senseval3	
	Method	**Correct/All**	Precision	**Correct/All**	**Precision**
1	IMSEMBED (On Wiki Corpus- Original)	3070/4328	70.9 (%)	2990/3944	75.8 (%)
2	IMS EMBED + Coeff (V1)	3068/4328	70.9	2989/3944	75.8
3	IMS EMBED + Coeff (V2)	3071/4328	71.0	2993/3944	75.9
4	IMS EMBED + Wcount (V1)	3064/4328	70.8	2997/3944	76.0
5	IMS EMBED + Wcount (V2)	3065/4328	70.8	2989/3944	75.8
6	IMS EMBED + Weighting	3079/4328	71.1	2995/3944	75.9
7	IMSEMBED + PCA (400)	3076/4328	71.1	2993/3944	75.9
8	IMSEMBED + PCA (400) + Coeff (V1)	3075/4328	71.0	2988/3944	75.8
9	IMSEMBED + PCA (400) + Coeff (V2)	3092/4328	71.4	2995/3944	75.9
10	IMS EMBED + PCA (400) + Wcount (V1)	3076/4328	71.1	2995/3944	75.9
11	IMS EMBED + PCA (400) + Wcount (V2)	3075/4328	71.0	2989/3944	75.8
12	IMSEMBED + PCA (400) + Weighting	3084/4328	71.3	2989/3944	75.8
13	IMSEMBED + PCA (400) + Coeff (V1) + Weighting	3088/4328	71.3	2991/3944	75.8
14	IMSEMBED + PCA (400) + Coeff (V2) + Weighting	3086/4328	71.3	2995/3944	75.9
15	IMSEMBED + PCA (400) + Wcount (V1) + Weighting	3086/4328	71.3	2998/3944	76.0
16	IMSEMBED + PCA (400) + Wcount (V2) + Weighting	3085/4328	71.3	2989/3944	75.8
17	IMSEMBED + Coeff (V1) + Weighting	3086/4328	71.3	2988/3944	75.8
18	IMSEMBED + Coeff (V2) + Weighting	3085/4328	71.3	2995/3944	75.9
19	IMSEMBED + Wcount (V1) + Weighting	3077/4328	71.1	2991/3944	75.8
20	IMSEMBED + Wcount (V2) + Weighting	3078/4328	71.1	2991/3944	75.8

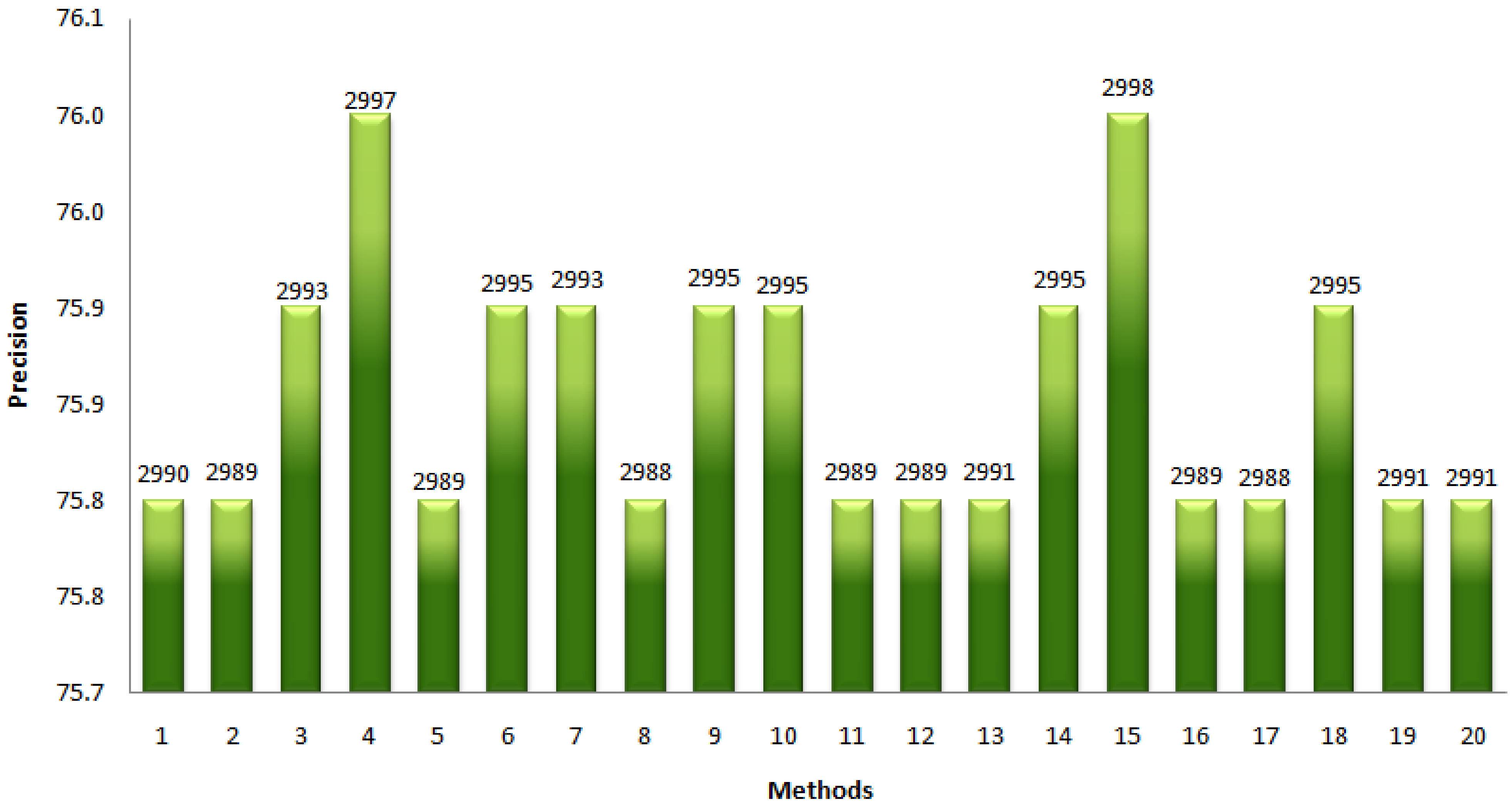
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Results of using the voting scheme:

	Senseval 2		Senseval 3	
Method	Correct	Precision (%)	Correct	Precision (%)
IMSE	3070	70.9	2990	75.8
IMSE + Voting	3071	71.0	3004	76.2

Conclusion

- Word embeddings information can be used in WSD task
- We reviewed A novel and simple unsupervised method to disambiguate words by deploying the trained word embeddings model of another language using only a bilingual dictionary.
- The main idea of this work is to use information provided by English-translated surrounding words to disambiguate Persian words using trained English word2vec model.

Conclusion (cont.)

- In the second part, we introduced four improvements to existing state-of-the-art supervised WSD approaches:
  - A new model for assigning vector coefficients
  - Applying a PCA dimensionality reduction process to find a better transformation of feature matrices
  - A new weighting scheme
  - A voting strategy to combine word embedding features extracted from different independent corpora.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

References:

- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity.
- Behzad Moradi and Ebrahim Ansari, “Unsupervised Word Sense Disambiguation using Word Embeddings” 2019 (under review).
- Sadi, M. F., Ansari, E. and Afsharchi, M., “Supervised Word Sense Disambiguation Using New Features Based on Word Embeddings,” 2019 (under review).

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Thanks.