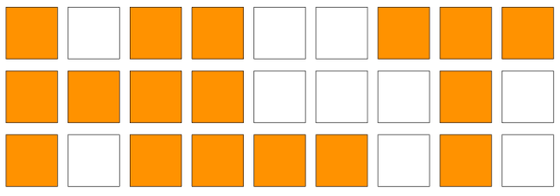
Word Embeddings Go to Italy: a Comparison of Models and Training Datasets

Giacomo Berardi,
Andrea Esuli and Diego Marcheggiani

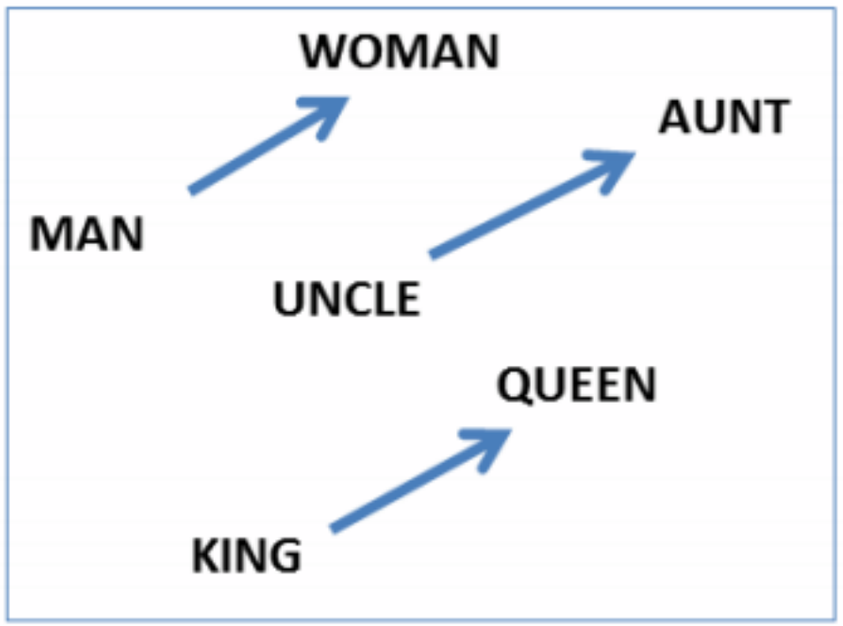


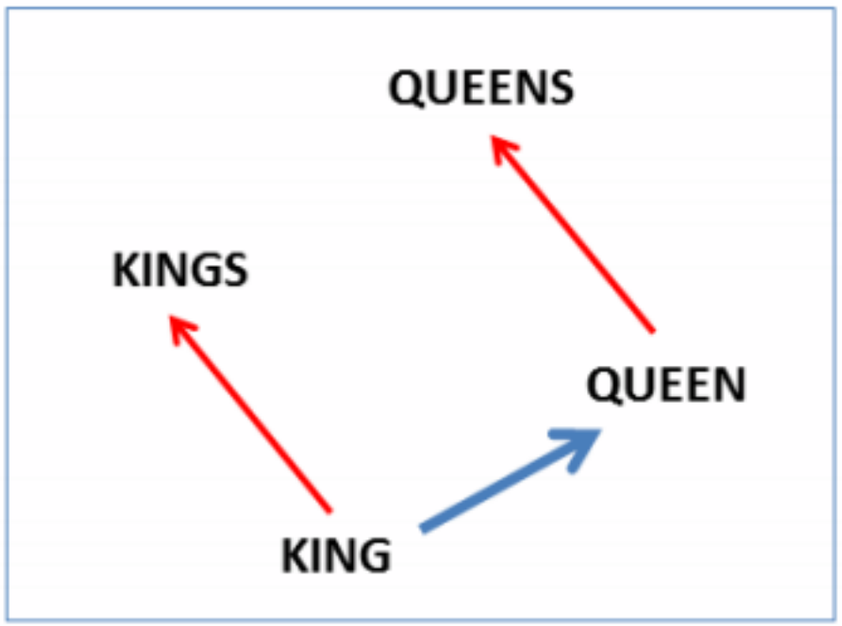
ISTI - Consiglio Nazionale
delle Ricerche, Pisa, Italy

IIR 2015
May 25, 2015

Introduction

- We present some preliminary results on the generation of word embeddings for the Italian language
- Semantic vector space model of language: representing each word with a real-valued vector





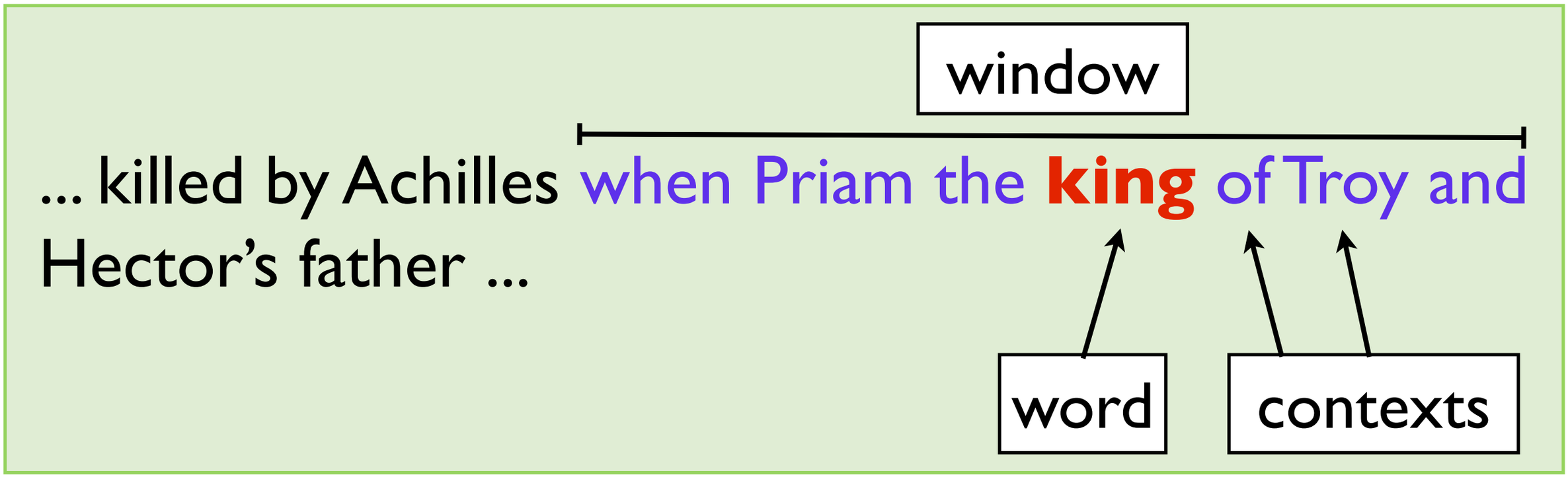
vect(“king”) - *vect*(“man”) + *vect*(“woman”) = *vect*(“queen”)

Examples in IR

- **Hierarchical Neural Language Models for Joint Representation of Streaming Documents and their Content** (WWW 2015)
  - Learn document and word contexts in the same model
- **Predicting The Next App That You Are Going To Use** (WSDM 2015)
  - App usage sessions are contexts from which features are generated
- **Fast and Space-Efficient Entity Linking in Queries** (WSDM 2015)
  - Relevance of a query term to an entity from their vector representation

Exploiting contexts

- Exploit word co-occurrences
- Shift a window on the sentences

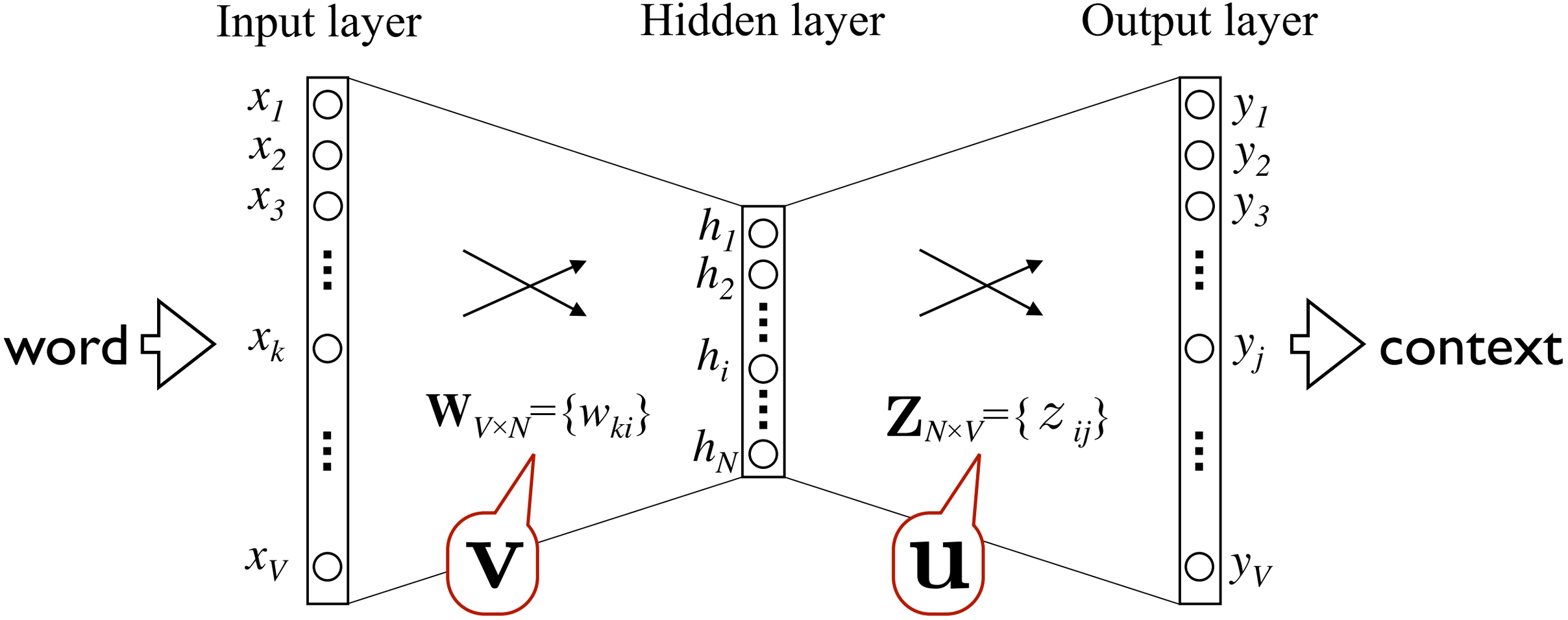


- We use two methods: **word2vec** and **GloVe**

word2vec skip-gram model

Learn $\mathbf{W}$ and $\mathbf{Z}$ so that the probability of having a context c, given a word i, is maximized

$$p(w_c|w_i) = \frac{\exp(\mathbf{u}_{w_c}^T \mathbf{v}_{w_i})}{\sum_{c'=1}^{C} \exp(\mathbf{u}_{w_{c'}}^T \mathbf{v}_{w_i})}$$



GloVe

- $P_{ij} = p(j|i)$ is the probability that word w_j appears in the context of word w_i
- The objective function:

$$F(\mathbf{v}_{w_i}, \mathbf{v}_{w_j}, \mathbf{u}_{w_c}) = P_{ic}/P_{jc}$$

- Cast the equation as a least squares problem (training through gradient descent). Minimize:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(\mathbf{v}_{w_i} \cdot \mathbf{u}_{w_j} + b_i + b_j - logX_{ij})^2$$

Datasets

	Documents	Sentences	Vocabulary	Vocabulary $tf \geq 5$	Tokens
WIKI	3,521,118	9,483,245	6,246,253	733,392	311,874,402
BOOKS	31,432	232,777,927	9,443,100	1,910,809	2,222,726,367

- WIKI's content aims at convey knowledge with an encyclopedic style, in which a simple but verbose language is preferred (average sentence length: 32)
- BOOKS's content mostly aims at entertaining the reader, using a rich and complex language, including dialogues and first-person speech (average sentence length: 9)

Test set

- Translation of the Google word analogy test for English
  - 19,791 questions, e.g., padre : madre = nonno : nonna
- The Italian comparative is very limited
- Superlative section has been translated to Italian absolute superlatives
- Participle of English has been mapped to the Italian gerund
- Plural-verbs section split it in two sections, one using the third person, and one using the first person
- Added four more sections: `present-remote-past-verbs (1st person)`, `masculine-feminine-singular`, `masculine-feminine-plural`, `regione-capoluogo`

Experimental settings

- Implementations:
  - word2vec: http://radimrehurek.com/gensim/
  - GloVe: http://nlp.stanford.edu/projects/glove/
- Window size: 10
- Vector size: 300
- Accuracy: ratio between the number correct answers and the total number of questions
- Wrong answers also the cases in which the words were not in the model

Results

	w2v-Skip-gram		GloVe	
	WIKI	BOOKS	WIKI	BOOKS
capital-common-countries	87.55%	84.78%	66.40%	61.07%
capital-world	63.86%	47.35%	22.74%	21.89%
currency	5.31%	3.58%	1.27%	0.58%
city-in-state	29.19%	23.47%	11.76%	15.04%
regione-capoluogo	41.23%	23.10%	23.10%	16.08%
family	58.01%	67.98%	51.44%	59.58%
accuracy on semantic	48.81%	38.56%	21.33%	21.54%
adjective-to-adverb	12.58%	17.74%	11.51%	14.52%
opposite	7.43%	27.54%	8.15%	25.91%
comparative	0.00%	8.33%	8.33%	8.33%
superlative (absolute)	8.72%	48.03%	16.83%	34.03%
present-participle (gerund)	55.02%	77.84%	51.89%	78.50%
nationality-adjective	77.36%	77.30%	68.17%	52.47%
past-tense	19.60%	61.36%	32.39%	63.64%
plural	40.79%	54.21%	31.67%	50.40%
plural-verbs (3rd person)	72.20%	85.73%	54.98%	74.29%
plural-verbs (1st person)	0.54%	45.05%	0.11%	30.32%
present-remote-past-verbs (1st person)	0.43%	34.41%	0.11%	18.60%
masculine-feminine-singular	35.71%	57.58%	33.12%	41.13%
masculine-feminine-plural	4.11%	29.00%	3.03%	15.37%
accuracy on syntactic	32.62%	54.56%	30.20%	44.60%
overall accuracy	39.91%	47.35%	26.21%	34.21%

Results

BOOKS	WIKI
scrivevo	
leggevo	desideravo
scrissi	suonavo
copiavo	figurarmi
riscrivevo	ascoltavo
rileggevo	innamorai
ricopiavo	rendessi

BOOKS	WIKI
batman	
superman	superman
supereroe	gotham
spiderman	catwoman
fumetti	batgirl
batgirl	joker
fumetto	batcaverna

Conclusions

- We tested two popular word embedding methods, trained on two Italian corpora
- We created a word analogy test set for the Italian language
- In the future we would:
  - Test other word embedding methods
  - Investigate the more appropriate datasets for specific tasks
  - Test word embeddings in practical applications

Thank You!

Dataset word vectors and the Italian word analogy test:
http://hlt.isti.cnr.it/wordembeddings

giacomo.berardi@isti.cnr.it