Word2vec and doc2vec (and how to evaluate them)

Word2vec

Two architectures:

- CBOW (Continuous Bag-of-words):

$$p(w_i | w_{i-h}, \ldots w_{i+h})$$

- Continuous Skip-gram:

$$p(w_{i-h}, \ldots w_{i+h} | w_i)$$

Two ways to avoid softmax:

- Negative sampling
- Hierarchical softmax

Open-source and fast: code.google.com/archive/p/word2vec/

Evaluation: word similarities

How do we test that *similar words* have *similar vectors*?

- Linguists know a lot about what is "similar".
- We can use *human judgements* for word pairs.
- Compare *Spearman's correlation* between two lists:

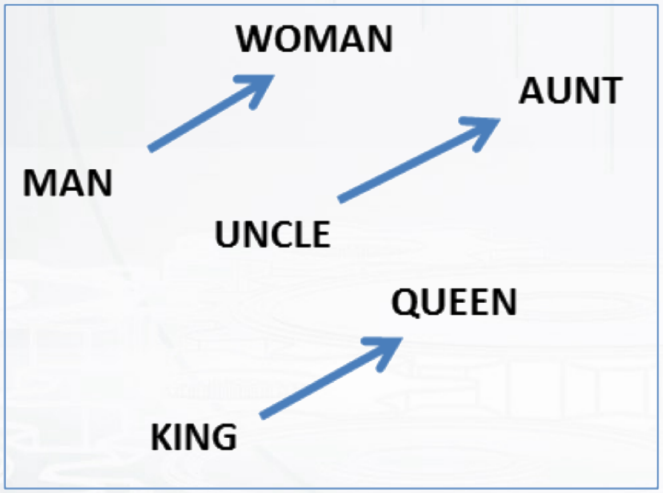
tiger	tiger	10.00
media	radio	7.42
tiger	cat	7.37
train	car	6.31
...	…	…

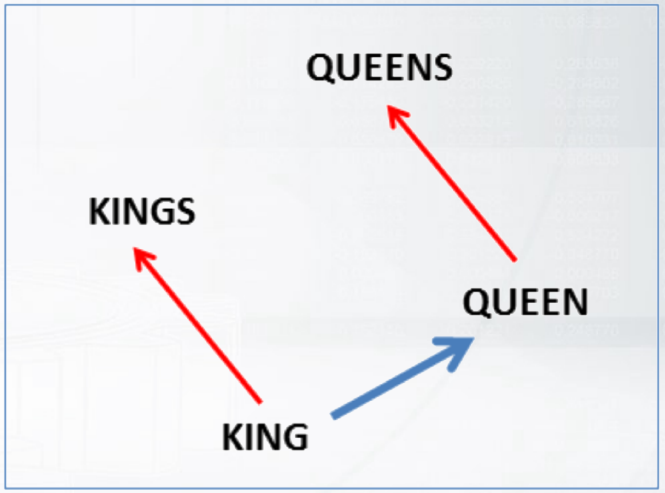
tiger	tiger	$\cos(\phi_u, \phi_v)$
media	radio	…
tiger	cat	…
train	car	…
...	…	…

Evaluation: word analogies

- In cognitive science well known as *relational similarity* (vs. *attributional similarity*).
- a : a' is as b : b' (man : woman is as king : ?)

$$\cos(b - a + a', x) \rightarrow \max_x$$





Gentner, D. Structure-mapping: A theoretical framework for analogy. Cognitive Science, 1983.
Mikolov et. al. Linguistic Regularities in Continuous Space Word Representations, 2013.

Word similarity task performance

- For word similarity task, count-based methods (PPMI, SVD) perform on par with predictive methods (GloVe, SGNS).

win	Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk
2	PPMI	.732	**.699**	.744	.654
	SVD	.772	.671	**.777**	.647
	SGNS	**.789**	.675	.773	**.661**
	GloVe	.720	.605	.728	.606
5	PPMI	.732	**.706**	.738	**.668**
	SVD	.764	.679	**.776**	.639
	SGNS	**.772**	.690	.772	.663
	GloVe	.745	.617	.746	.631

`win` is the width of the window for co-occurrences collection.

Levy et. al. Improving distributional similarity with lessons learned from word embeddings, 2015.

Word analogy task performance

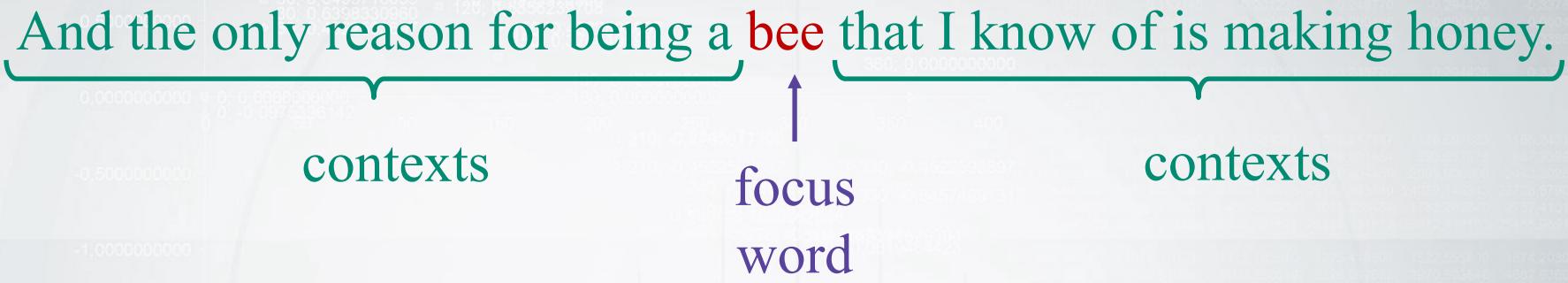
- Word analogy task is solved with 70% average accuracy.

win	Method	Google Add / Mul	MSR Add / Mul
2	PPMI	.552 / .677	.306 / .535
	SVD	.554 / .591	.408 / .468
	SGNS	.676 / **.689**	.617 / **.644**
	GloVe	.649 / .666	.540 / .591
5	PPMI	.518 / .649	.277 / .467
	SVD	.532 / .569	.369 / .424
	SGNS	.692 / **.714**	.605 / **.645**
	GloVe	.700 / .712	.541 / .599

`Add` is the way of analogy solving that we discussed. `Mull` is a modification.

Levy et. al. Improving distributional similarity with lessons learned from word embeddings, 2015.

Paragraph2vec aka doc2vec



DM (Distributed Memory):

$$p(w_i | w_{i-h}, \ldots w_{i+h}, d)$$

DBOW (Distributed Bag Of Words):

$$p(w_{i-h}, \ldots w_{i+h} | d)$$

Evaluation: document similarities

How do we test that *similar documents* have *similar vectors*?

- ArXiv triplets: paper A, similar paper B, dissimilar paper C
- Measure the accuracy of guessing the dissimilar paper

http://arxiv.org/pdf/1206.5743	http://arxiv.org/pdf/cond-mat/0403258	http://arxiv.org/pdf/1408.0189
http://arxiv.org/pdf/1209.0268	http://arxiv.org/pdf/1307.7598	http://arxiv.org/pdf/math/0504051
http://arxiv.org/pdf/hep-ph/9908436	http://arxiv.org/pdf/nucl-th/9707019	http://arxiv.org/pdf/1112.3014
http://arxiv.org/pdf/1111.2905	http://arxiv.org/pdf/1303.2538	http://arxiv.org/pdf/1109.1922
http://arxiv.org/pdf/nucl-ex/0112013	http://arxiv.org/pdf/physics/9704013	http://arxiv.org/pdf/1408.4595
http://arxiv.org/pdf/0709.3419	http://arxiv.org/pdf/quant-ph/0611134	http://arxiv.org/pdf/0902.0616
http://arxiv.org/pdf/hep-th/9609148	http://arxiv.org/pdf/solv-int/9710009	http://arxiv.org/pdf/astro-ph/0508060

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015.

Evaluation: document similarities

Integral formula of Minkowski type and new characterization of the Wulff shape

Yijun He * Haizhong Li †

Abstract

Given a positive function F on S^n which satisfies a convexity condition, we introduce the r-th anisotropic mean curvature M_r for hypersurfaces in $\mathbb{R}^{n+1}$ which is a generalization of the usual r-th mean curvature H_r. We get integral formulas of Minkowski type for compact hypersurfaces in R^{n+1}. We give some new characterizations of the Wulff shape by use of our integral formulas of Minkowski type, in case $F = 1$ which reduces to some well-known results.

2000 Mathematics Subject Classification: Primary 53C42, 53A30; Secondary 53B25.

Key words and phrases: Wulff shape, F-Weingarten operator, anisotropic principal curvature, r-th anisotropic mean curvature, integral formula of Minkowski type.

xiv.org/pdf/1408.0189

xiv.org/pdf/math/0504051

xiv.org/pdf/1112.3014

xiv.org/pdf/1109.1922

xiv.org/pdf/1408.4595

xiv.org/pdf/0902.0616

xiv.org/pdf/astro-ph/0508060

Evaluation: document similarities

Integral formula of Minkowski type and new characterization of the Wulff shape

COMPLEX CURVES IN ALMOST-COMPLEX MANIFOLDS AND MEROMORPHIC HULLS

Sergei IVASHKOVICH – Vsevolod SHEVCHISHIN

Preface

Chapter I. Local Properties of Complex Curves.

Lecture 1. Complex Curves in Almost-Complex Manifolds. . . . pp. 1–12
1.1. Almost-Complex Manifolds, Hermitian Metrics, Associated (1,1)-Forms. 1.2. Existence of Calibrating and Tame Structures. 1.3. Almost-Complex Submanifold, Complex Curves, Energy and Area. 1.4. Symplectic Surfaces. 1.5. Adjunction Formula for Immersed Symplectic Surfaces.

Evaluation: document similarities

Accepted for publication in *Solar Physics*, waiting for the authoritative version and a DOI which will be available at http://www.springerlink.com/content/0038-0938

Time-dependent Stochastic Modeling of Solar Active Region Energy

M. Kanazir and M. S. Wheatland[1]

Received: 7 July 2010 / Accepted: 31 July 2010 / Published online: ••••••••••

Abstract A time-dependent model for the energy of a flaring solar active region

Lecture 1. Complex Curves in Almost-Complex Manifolds. . . . pp. 1–12
1.1. Almost-Complex Manifolds, Hermitian Metrics, Associated (1,1)-Forms. 1.2. Existence of Calibrating and Tame Structures. 1.3. Almost-Complex Submanifold, Complex Curves, Energy and Area. 1.4. Symplectic Surfaces. 1.5. Adjunction Formula for Immersed Symplectic Surfaces.

Resume

Methods:

- *word2vec:* SGNS, CBOW, …
- *doc2vec:* DBOW, DM, …
- Python library for both: https://radimrehurek.com/gensim/

Evaluation:

- Word similarity and analogy
- Document similarity
- *Interpretability of the components*
- *Geometry of the embeddings space*

Count-based and predictive approaches are not so different!