Unsupervised Word Usage Similarity in Social Media Texts

Spandana Gella and Paul Cook and Bo Han

Department of Computing and Information Systems
The University of Melbourne

Social media — Twitter

- Huge volume of user generated text
  - Applications: trend analysis, event detection, natural disaster response co-ordination
- Short, noisy text: Challenging for traditional NLP
- Little work to-date on lexical semantics on Twitter

This paper: Lexical semantic interpretation in tweets based on word usage similarity

Word sense disambiguation (WSD)

Given a word in context, select the best-fitting "sense" from a sense inventory

- *ne1 headin to blue boyz footy match this weekend? #iamcarlton*
  - game in which players or teams compete against each other
- *I think they are a perfect match! #cute #xoxo*
  - something that resembles or harmonizes; "that tie makes a good match with your jacket"
  - a pair of people who live together; "a married couple from Chicago"

Word sense disambiguation (WSD)

Given a word in context, select the best-fitting "sense" from a sense inventory

- *ne1 headin to blue boyz footy match this weekend? #iamcarlton*
  - game in which players or teams compete against each other
- *I think they are a perfect match! #cute #xoxo*
  - something that resembles or harmonizes; "that tie makes a good match with your jacket"
  - a pair of people who live together; "a married couple from Chicago"

Issues with WSD

- Choice of sense inventory
  - *match*: WordNet 9 senses, Macmillan 4 senses
- No sense tagged resources for social media data
- Cannot capture novel usage patterns
- Assumes a single sense per usage
- Challenges over social media: short, noisy, non-standard syntax

Solution? An alternative representation of meaning in context

Usage similarity (Usim)

- The manual task of rating the similarity of a pair of usages of a word (SPair) [Erk et al., 2009].
- Similarity on a graded scale (1 – 5)
- No more senses; independent of sense inventory
- Novel usages: Rate similarity to established usages

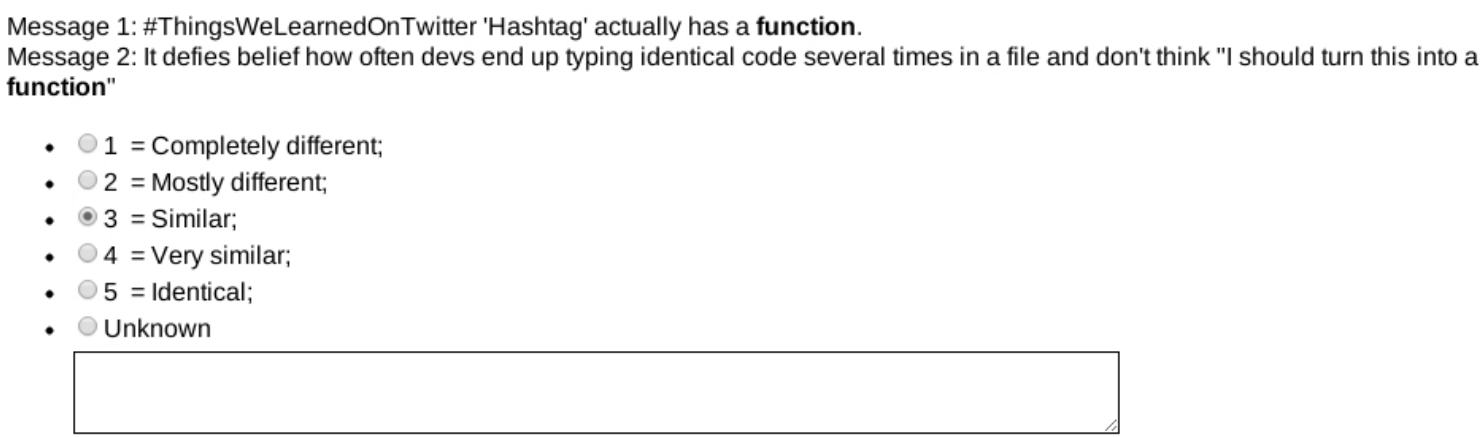
SPair example (annotators' judgement: 3.2)

- *Setting goals for myself this year, figured if it's on paper I'll be more inspired to work harder.*
- *This is very unsmart of me to get tipsy and then have to go home and write a paper.*

Gold standard — Usim-tweet annotation

- 10 nouns: *bar*, *charge*, *execution*, *field*, *figure*, *function*, *investigator*, *match*, *paper*, *post*
- 55 pairs of tweets (SPairs) were annotated per lemma (sampled from Twitter Streaming API)
- Amazon Mechanical Turk annotation

Sample Annotation:


Message 1: #ThingsWeLearnedOnTwitter 'Hashtag' actually has a **function**.
Message 2: It defies belief how often devs end up typing identical code several times in a file and don't think "I should turn this into a **function**"

- 1 = Completely different;
- 2 = Mostly different;
- 3 = Similar;
- 4 = Very similar;
- 5 = Identical;
- Unknown

Background corpora

- ORIGINAL: Tweets from Streaming API containing target word as noun
  - *gotta 3 page* **paper** *due tomorrow haven start #procrastination*
- EXPANDED: ORIGINAL + document expansion based on medium-frequency hashtags
  - *research* **paper** *due tomorrow 2 page 3 #procrastination*
  - *insert figure equations lab report less time word #physicsmajor #procrastination ...*
- RANDEXPANDED: ORIGINAL + extra tweets containing target

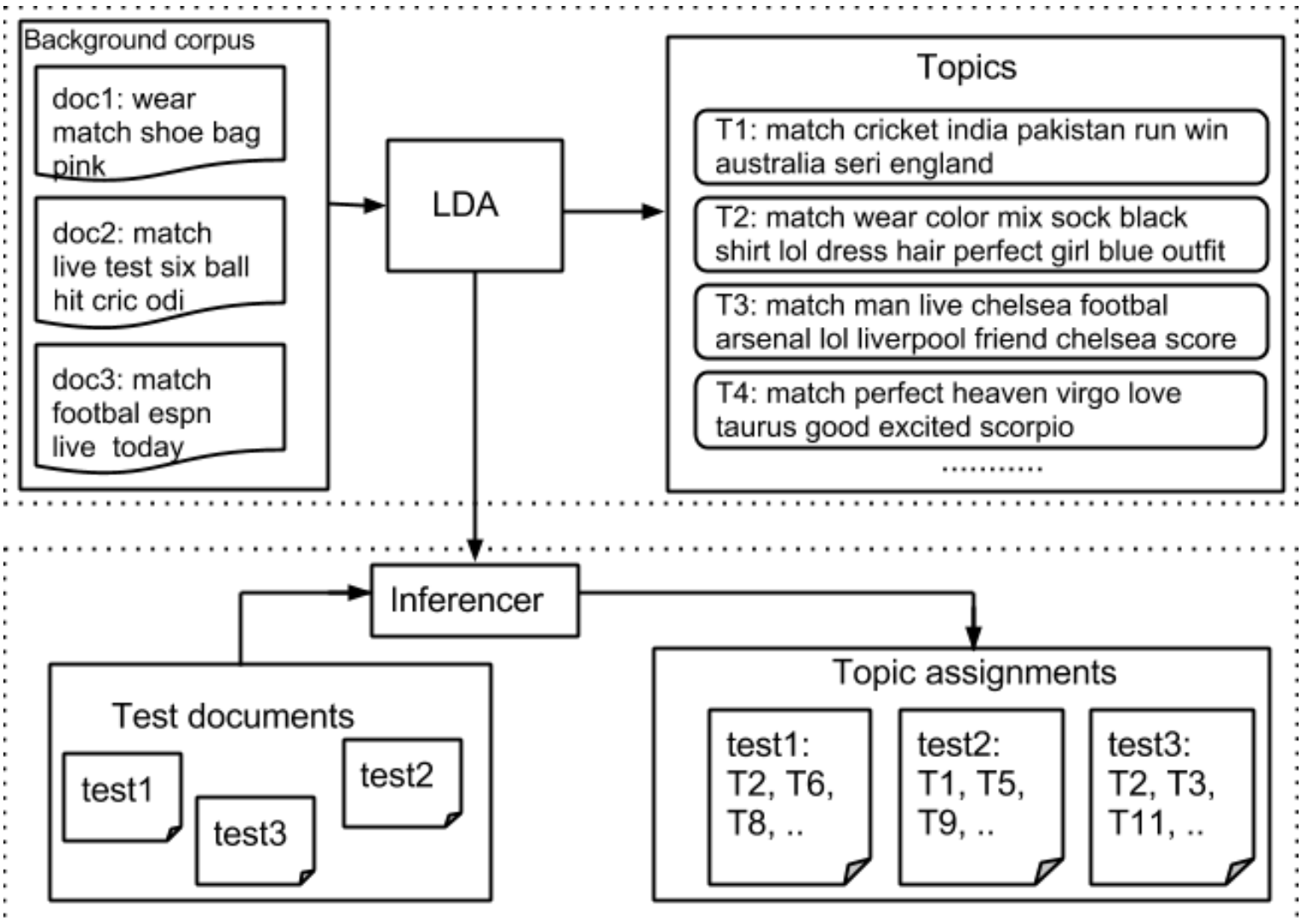
Model Usim for Twitter

- No large annotated training resources: unsupervised
- No parser: bag-of-words
- Methods?
  - Vector space model (VSM)
  - Topic models (LDA) [Lui et al., 2012]
  - Weighted Textual Matrix Factorization (WTMF)
- 1 model per target word

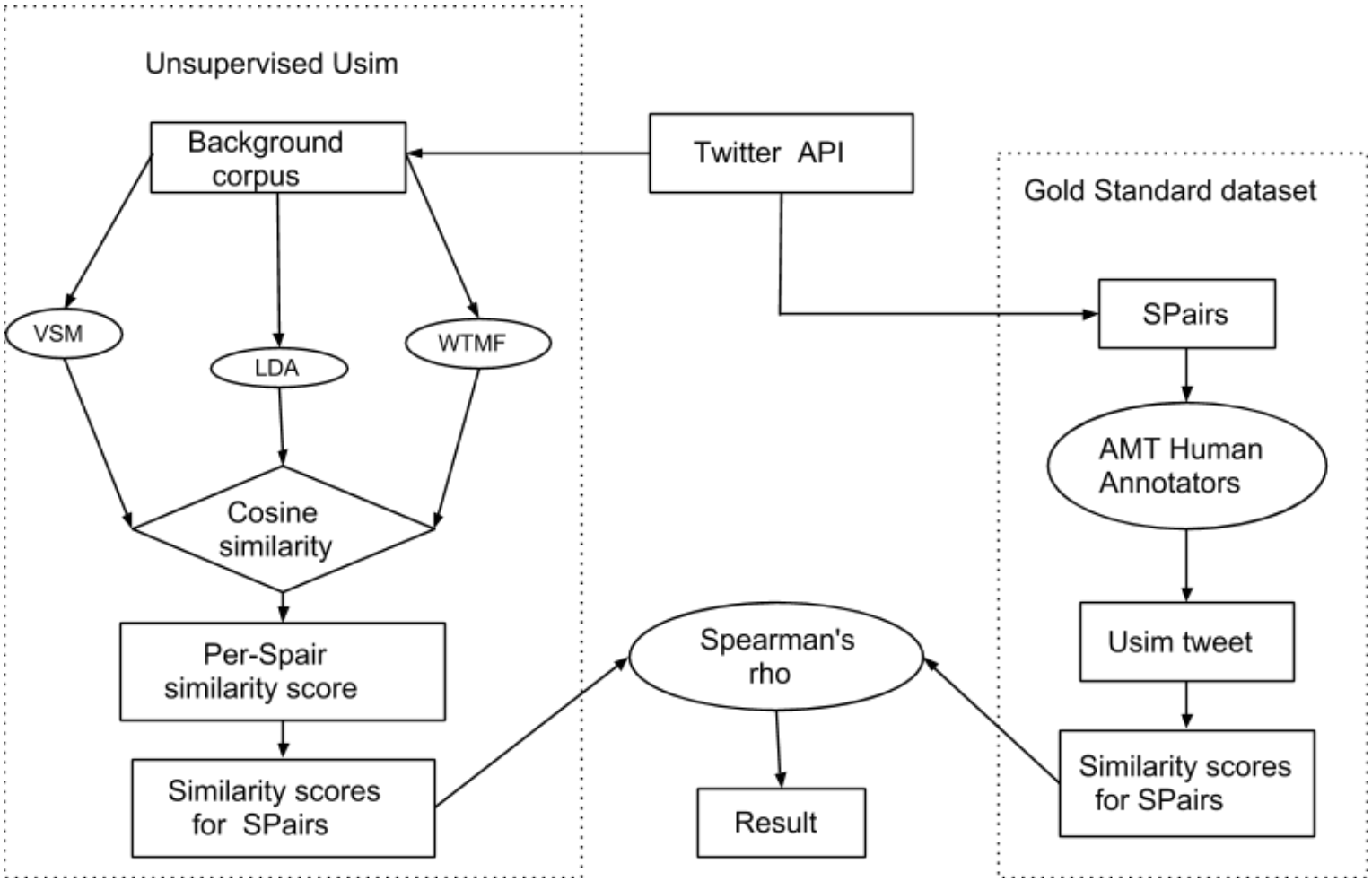
Methods

- VSM (Baseline): Second order co-occurrence
- LDA (Our approach): Represent documents as topic distribution vectors
- WTMF (Benchmark): Consider information from "missing" words related to the latent vector profile
  - State-of-the-art on a similar task

Topic modeling — LDA



Method overview



Results

Model	ORIGINAL	EXPANDED	RANDEXPANDED
Baseline	0.09	0.08	0.09
WTMF (d)	0.03 (8)	0.10 (20)	0.09 (5)
LDA (T)	**0.20 (8)**	**0.29 (5)**	**0.18 (20)**

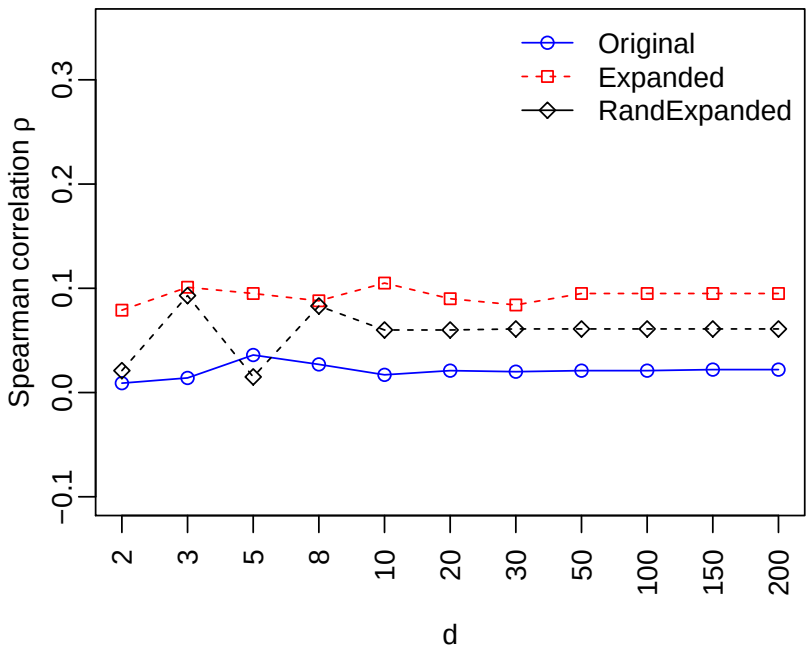
Spearman's rank correlation (ρ) for each method based on each background corpus

Results on EXPANDED

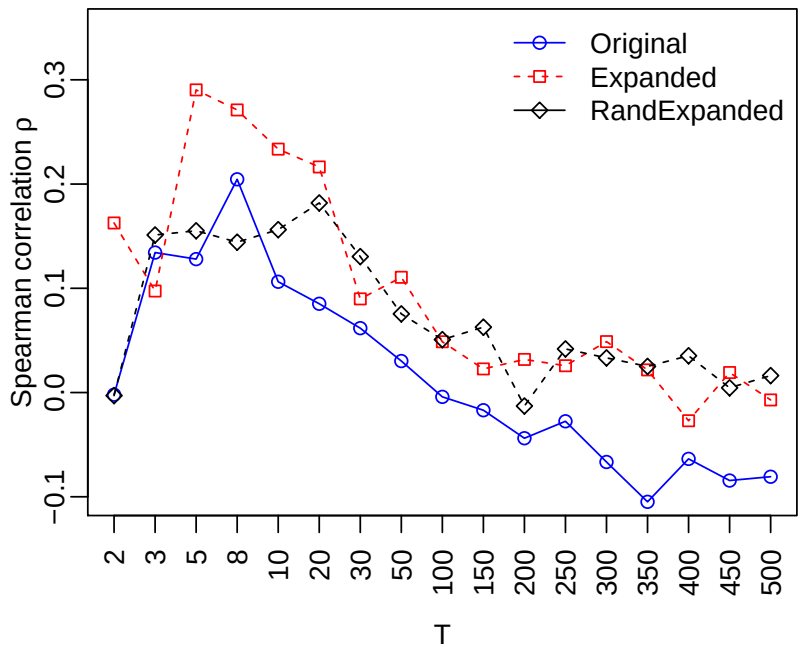
	Best-T		5-topics
Lemma	ρ	T	ρ
bar	**0.35**	50	0.1
charge	**0.33**	20	−0.08
execution	**0.58**	5	**0.58**
field	**0.53**	10	0.32
figure	0.24	250	0.14
function	**0.40**	10	**0.27**
investigator	**0.50**	5	**0.50**
match	**0.45**	5	**0.45**
paper	0.32	30	0.22
post	0.2	30	−0.01
Overall	**0.29**	5	**0.29**

ρ values that are significant ($p > 0.05$) are shown in bold

Results varying d and T



(a) WTMF: ρ versus dimensions (d)



(b) LDA: ρ versus topics (T)

Summary

- Computationally modeled Usim over social media
- LDA approach out-performed a baseline and benchmark
- Hashtag based document expansion improved performance of LDA and benchmark
- Gold-standard dataset Usim-tweet will be made available
- Future work
  - Alternative document expansion (e.g., author based)
  - Context representation: POS, positional word features, etc.
  - Non-parametric topic modelling (e.g., HDP)

Thanks

Bibliography

Erk, K., McCarthy, D., and Gaylord, N. (2009).

Investigations on word senses and word usages.

In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, pages 10–18, Singapore.

Lui, M., Baldwin, T., and McCarthy, D. (2012).

Unsupervised estimation of word usage similarity.

In *Proceedings of the Australasian Language Technology Workshop 2012 (ALTW 2012)*, pages 33–41, Dunedin, New Zealand.