Understanding Social Tags: Relation Extraction and Tag Annotation

Presentation at NLP@UoL, Mar 23, 2018

Hang Dong

Supervisors: Wei Wang, Frans Coenen, Kaizhu Huang

Acknowledgement to the all the figures and tables used from (García-Silva et. al , 2012; Bahdanau, Cho & Benjio, 2015; Yang et al., 2016; Li et al., 2016)

Introduction

- Hang Dong, http://www.csc.liv.ac.uk/~hang/
- Third (2.5) Year PhD student,
- UoL (Based at Xi'an Jiaotong-Liverpool University)
- Research visit @UoL from 20 Feb 2018 to 21 May 2018.

- MSc Information Systems, Information School, University of Sheffield, 2013-2014.
- BMgt Library Sciences, Wuhan University, Wuhan, China, 2009-2013

Overview

- Relation Extraction: Automatic Taxonomy Generation from Social Tagging Data to Enrich Knowledge Bases
    - Feature extracted from probabilistic topic analysis of tags.

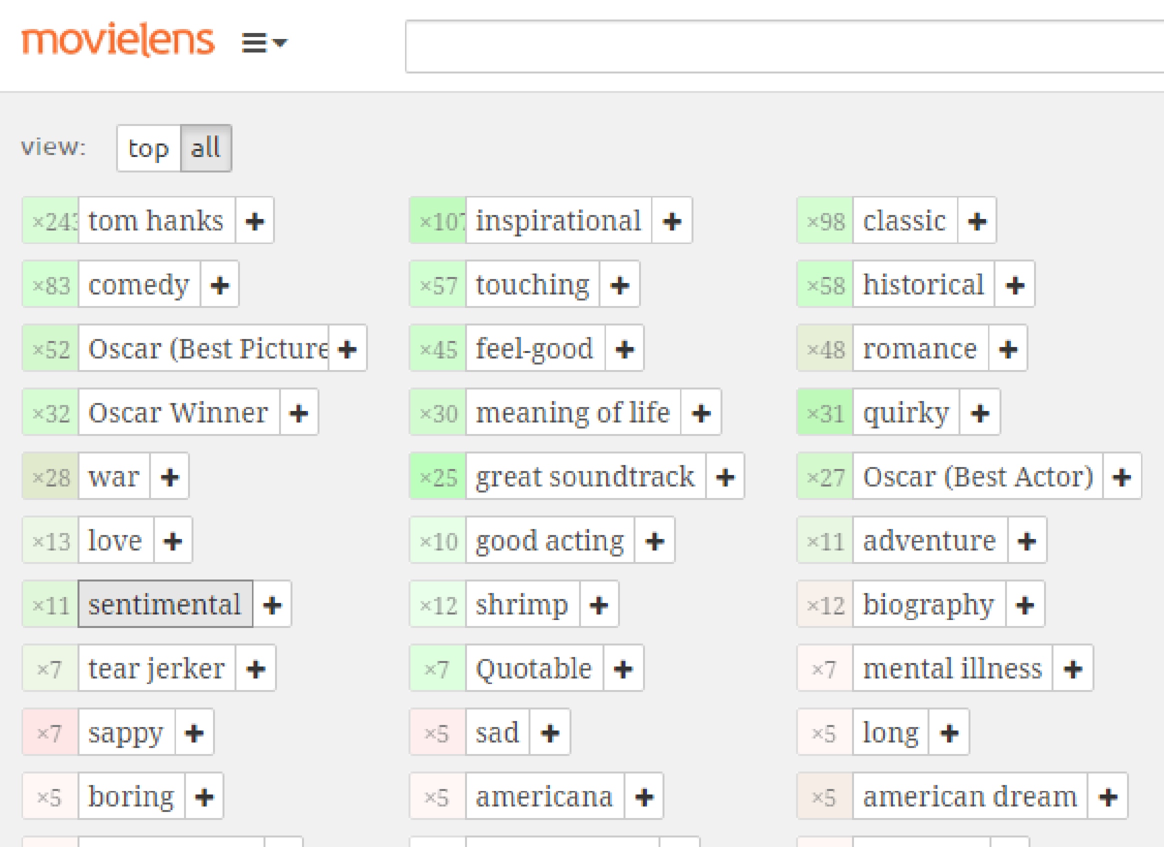
- Tag Annotation: Sequence Modelling for Tag Annotation / Recommendation
    - Focus on attention mechanisms for tag annotation.

Motivation – Organising social tags semantically

- Social tagging: Users share a resource – create short text description – terminology of a social group / a domain

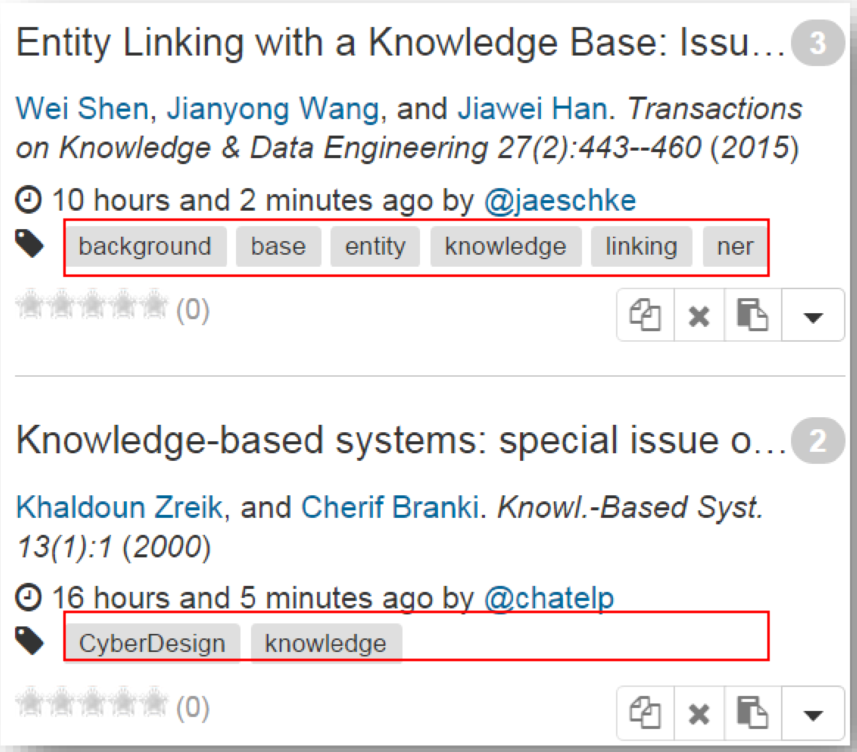
- "Folksonomy [social tags] is the result of personal free tagging of pages and objects for one's own retrieval" (Thomas Vander Wal, 2007)

- Noisy and ambiguous, thus not useful to support information retrieval and recommendation.



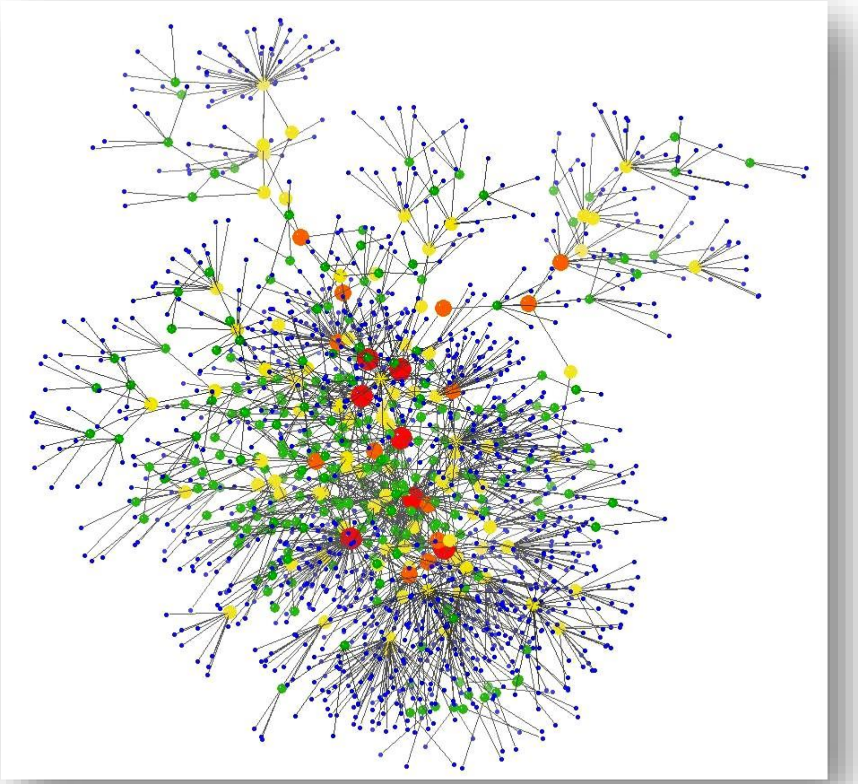
Social tags for movie "Forrest Gump" in MovieLens
https://movielens.org/movies/356

Research aim: from academic social data to knowledge



http://www.bibsonomy.org/tag/knowledge

Researcher generated data (user-tag-resource-date)



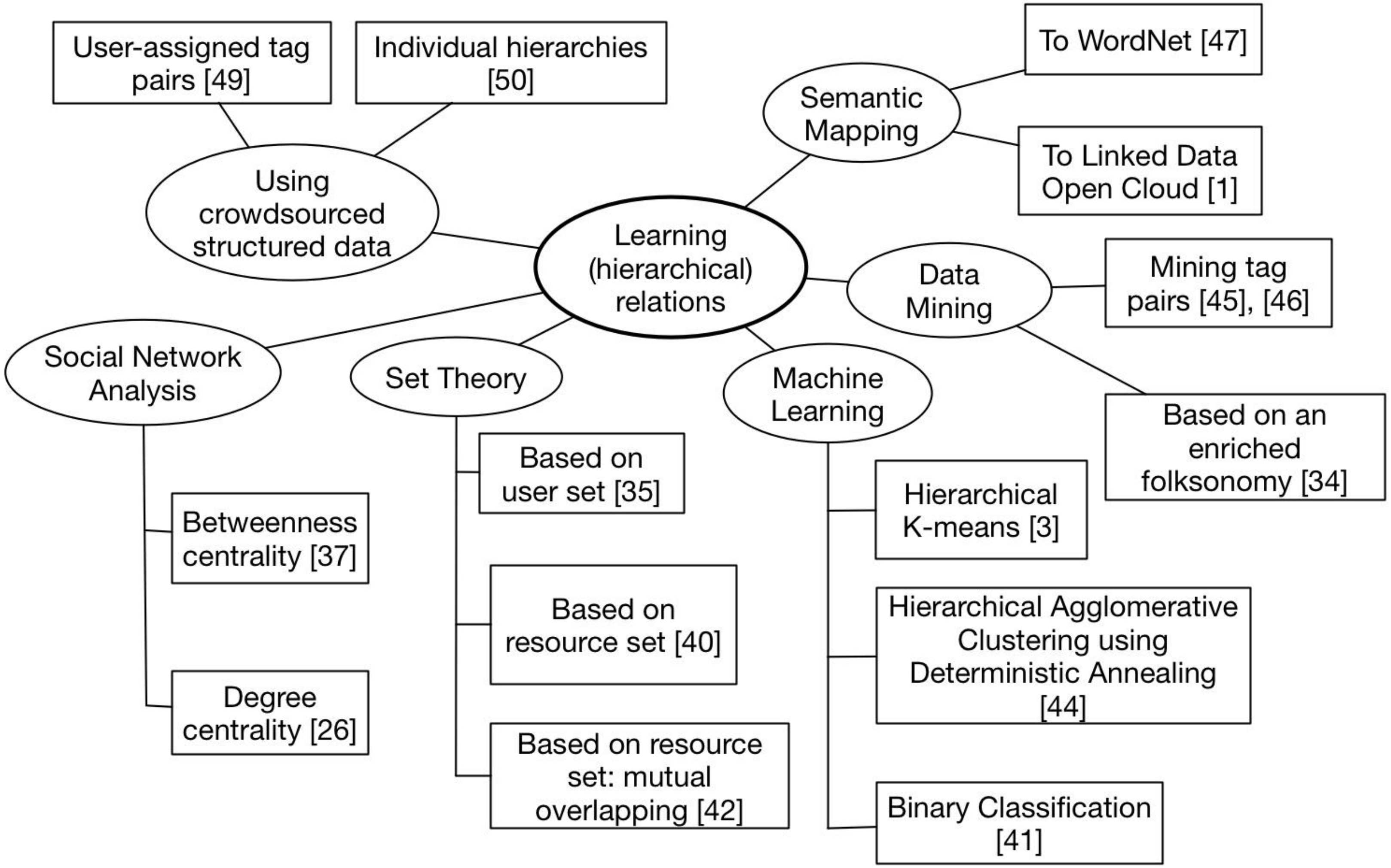
http://www.micheltriana.com/blog/2012/01/20/ontology-what

Useful and evolving knowledge structure

Challenges

- Distinct from text corpora: Lack of context information
  - Pattern-based approaches (Hearst patterns) do not work.
- Noise in data
- Sparsity in data

Relation extraction
Learning (hierarchical) relations from social tagging data



H. Dong, W. Wang and H.-N. Liang, "Learning Structured Knowledge from Social Tagging Data: A Critical Review of Methods and Techniques," *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, Chengdu, 2015, pp. 307-314.

Types and issues of current methods

- **Heuristics based methods** (set inclusion, graph centrality and association rule) are based on co-occurrence, does not formally define semantic relations (Garc'ia-Silva et al., 2012).

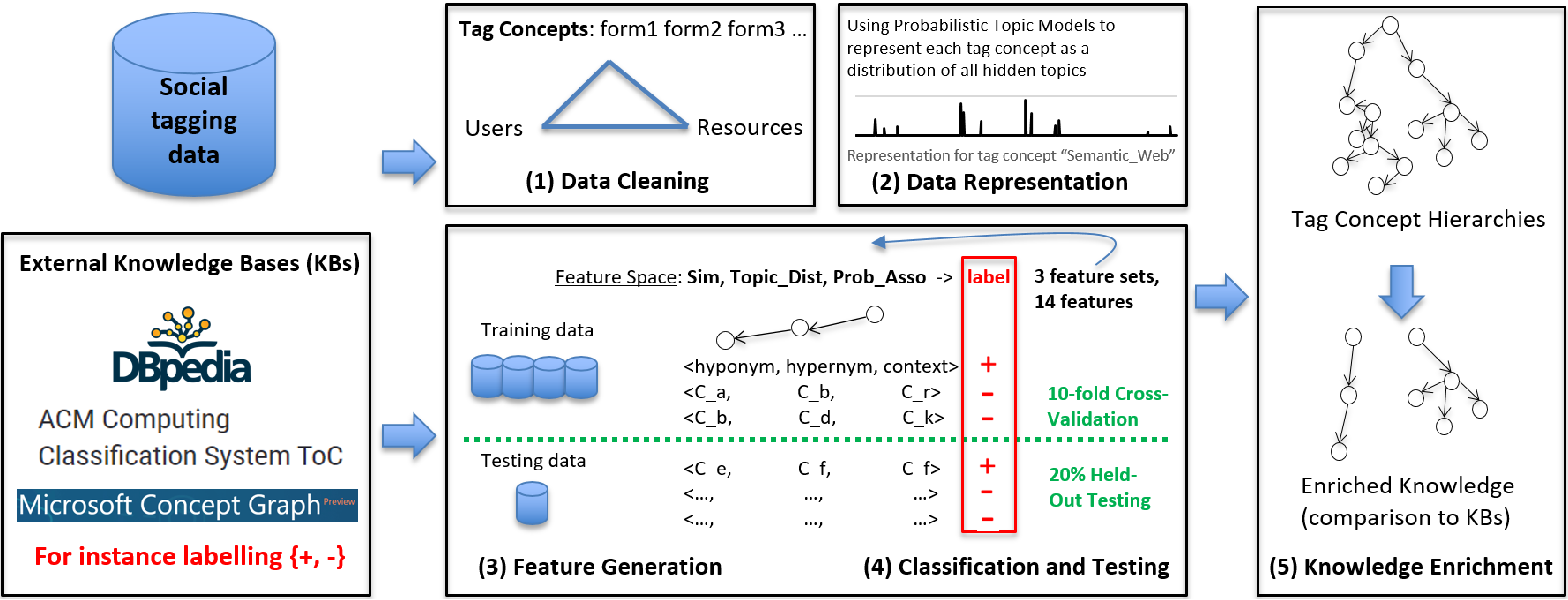
- **Semantic grounding methods** (matching tags to lexical resources) suffer from the low coverage of words and senses in the relatively static lexical resources (Andrews & Pane, 2013; Chen, Feng & Liu, 2014).

- **Machine learning methods**: (i) unsupervised methods could not discriminate among subordinate, related and parallel relations (Zhou et al., 2007); (ii) supervised methods so far based on data co-occurrence features (Rego, Marinho & Pires, 2015).

- We proposed a new supervised method, binary classification founded on a set of assumptions using probabilistic topic models.

Supervised learning based on Probabilistic Topic Modeling

Binary classification: input two tag concepts with a context tag, output whether they have a hierarchical relation. There are 14 features.



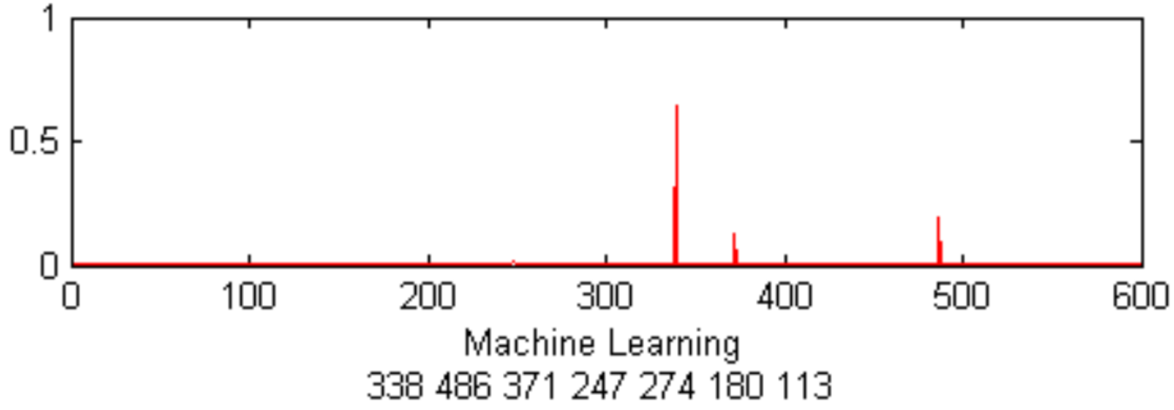
Data Representation

- We used a unsupervised approach **Probabilistic Topic Model**, Latent Dirichlet Allocation, to infer the hidden topics in the Bag-of-Tags used to annotate resources. Then we represented each tag as a probability on the hidden topics, reduced dimensionality of the vector space.
- Input: Bag-of-tags (resources) as documents
- Output: p(word | topic), p(topic | document)

$$v(C_a) = \{p(\mathbf{z}_i|C_a)\}_{i=1}^{\|\mathbf{z}\|} \quad (1)$$

$$p(z|C_a) \propto p(C_a|z) * p(z) \quad (2)$$

$$p(z) = \frac{N_z}{N} \quad (3)$$



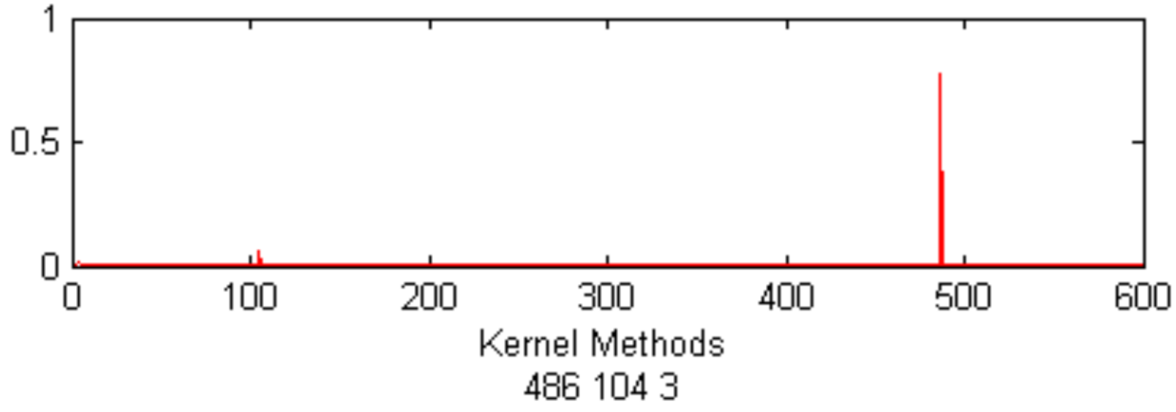


TABLE VII
TAG TOPICS LEARNED USING LATENT DIRICHLET ALLOCATION (LDA)
($T = 600$, ALPHA $= 50/600$, BETA $= 0.01$)

Topic	Most probable 5 tag concepts
62	search web web_search semantic_search social_search
154	cell calcium membrane channel animal
159	language perception speech tone production
231	game game_theory learning theory haifa_games_course
369	child male female cerebral human

Assumptions and Feature Generation

- Assumption 1 (Topical Similarity) For two tag concepts, they must be similar enough, in terms of a similarity measure, to have a hierarchical relation.

TABLE II
SIMILARITY AND DIVERGENCE RELATED FEATURES

Features	Description
Cos_sim	The cosine similarity of two topic distribution vectors
KL_Div1	The Kullback-Leibler Divergence from C_a to C_b
KL_Div2	The Kullback-Leibler Divergence from C_b to C_a
Gen_Jaccard	The generalised Jaccard Index of two topic distribution vectors

For the generalised Jaccard Index,

$$\mathbf{J}(\mathbf{v}(\mathbf{C_a}), \mathbf{v}(\mathbf{C_b})) = \frac{\sum_i \min(v(C_a)_i, v(C_b)_i)}{\sum_i \max(v(C_a)_i, v(C_b)_i))} \qquad (5)$$

- Assumption 2 (Topic Distribution): a tag more evenly distributed on several topics may have a sense more general than a tag distributed on fewer topics.

TABLE III
TOPIC DISTRIBUTION RELATED FEATURES

Features	Description
diff_num_sig	Difference of the number of significant topics
overlapping	Number of overlapping significant topics
diff_max	Difference of the maximum elements in two tag vectors
diff_aver_sig	Difference of the average probability of significant topics

$$\mathbf{z}_a^{sig} = \{z \mid z \in \mathbf{z} \text{ and } p(z|C_a) \geqslant p\} \quad (4)$$

$$\begin{aligned} \text{diff\_aver\_sig}(C_a, C_b) &= \text{Aver}(\mathbf{z}_a^{sig}) - \text{Aver}(\mathbf{z}_b^{sig}) \\ &= \frac{\sum(\mathbf{z}_a^{sig})}{\|\mathbf{z}_a^{sig}\|} - \frac{\sum(\mathbf{z}_b^{sig})}{\|\mathbf{z}_b^{sig}\|} \end{aligned} \quad (6)$$

$\mathbf{z}_a^{sig}$ is the significant topic set for the concept C_a.

$\mathbf{z}$ is the whole topic set.

p is a probability threshold.

- Assumption 3 (Probabilistic Topical Association) For two tag concepts, if they have strong conditional probability marginalised on topics, they are more likely to have a hierarchical relation.

TABLE IV
PROBABILISTIC ASSOCIATION FEATURES

Features	Description
$p(C_a\|C_b)$	The probabilistic association of C_a given C_b
$p(C_b\|C_a)$	The probabilistic association of C_b given C_a
$p(C_a, C_b)$	The joint probability of C_a and C_b
$p(C_a\|C_b, R_{a,b})$	The probabilistic association of C_a given C_b and the common root concept $R_{a,b}$
$p(C_b\|C_a, R_{a,b})$	The probabilistic association of C_b given C_a and the common root concept $R_{a,b}$
$p(C_a, C_b\|R_{a,b})$	The joint probability of C_a and C_b given the common root concept $R_{a,b}$

$$\begin{aligned} p(C_a|C_b) &= \sum_{z\in\mathbf{z}} p(C_a|z, C_b)p(z|C_b) \\ &= \sum_{z\in\mathbf{z}} p(C_a|z)p(z|C_b) \end{aligned} \tag{7}$$

$$\begin{aligned} p(C_a, C_b) &= p(C_a|C_b)p(C_b) \\ &= p(C_a|C_b)\sum_{z\in\mathbf{z}} p(C_b|z)p(z) \end{aligned} \tag{9}$$

$$\begin{aligned} p(C_a|C_b, R_{a,b}) &= \sum_{z\in\mathbf{z}} p(C_a|z, C_b, R_{a,b})p(z|C_b, R_{a,b}) \\ &= \sum_{z\in\mathbf{z}} p(C_a|z)p(z|C_b, R_{a,b}) \\ &= \sum_{z\in\mathbf{z}} p(C_a|z)\frac{p(C_b, R_{a,b}|z)p(z)}{p(C_b, R_{a,b})} \\ &= \sum_{z\in\mathbf{z}} \frac{p(C_a|z)p(C_b|z)p(R_{a,b}|z)p(z)}{p(C_b, R_{a,b})} \end{aligned} \tag{8}$$

Hierarchy Generation Algorithm

- After we trained the model, we propose a greedy-search hierarchy generation algorithm to predict concept hierarchies from social tags.

- The algorithm has some characteristics:
    - Progressively predicts the hierarchy from top to down from a user specified root concept.
    - Generates a mono-hierarchy (a tree), each concept has only one hypernym (broader concept).
    - Prune the tree by keeping the relations with higher confidence score from the classification model.

Algorithm 1: Hierarchy Generation Algorithm: a heuristic-based greedy algorithm to learn and prune relations layer by layer to learn a standard monohierarchy.

Require: h, $root$, $context$, generateCand(), $Criteria$, I_i, generateFeature(I_i), predict(h, x_i), size(), TH_s

Ensure: G, an induced taxonomy as a directed graph.

Initialise G,G_{curr},G_{next};
L = generateCand($root$, $context$, $Criteria$);
for *each node in* L **do**
    form input instance set $I_i = <node, root, context\_root>$;
    x_i = generateFeature(I_i);
    y'_i = predict(h,x_i);
    if $y'_i > 0$ **then**
        $G_{curr} \leftarrow G_{curr} \cup < node, root >$;
    end
end
Remove all new established *node* in G_{curr} from L;
$G \leftarrow G \cup G_{curr}$;
while $size(L) > TH_s$ **do**
    G_{next} = learnNextLayer(G_{curr}, h, L, $Criteria$); % See Algorithm $\boxed{2}$
    $G_{curr} = G_{next}$;
    Remove all new established *node* in G_{curr} from L;
    $G \leftarrow G \cup G_{curr}$;
end

Input: a tag as root, and a tag as context
Output: Hierarchy

- Generate concept candidates for the hierarchy
- Do
  - Generate layer 1
  - Generate layer 2
  - Generate layer 3
  - …
  - Generate layer *n*
- Until not enough candidates

Evaluation - Dataset

- Social tagging data: Bibsonomy, 283858 tags, 11103 users, 868015 resources
- External Knowledge Bases (EKBs):
  - (i) DBpedia, (ii) Microsoft Concept Graph (MCG) and (iii) ACM Computing Classification System (CCS).
- After automatic labeling to the three EKBs:
  - 14535 instances (4965 positive instances, 4785 reversed negative instances, 4785 random negative instances.)
- Positive : Negative = 1:1.93

Table 7: Statistics of EKBs and overlapping to Bibsonomy

	Concepts	Subsumption relations	Concept overlapping to Bibsonomy	Release Date
DBpedia	1316674	2706685	2191	2015-10
MCG	1483135	2844951	6030	2016-09
ACM	9060	2390	691	2012
Bibsonomy	7458	-	-	2015-07

Data Cleaning and Concept Extraction

Using inter-subjectivity (user frequency) and edited distance to group word forms.

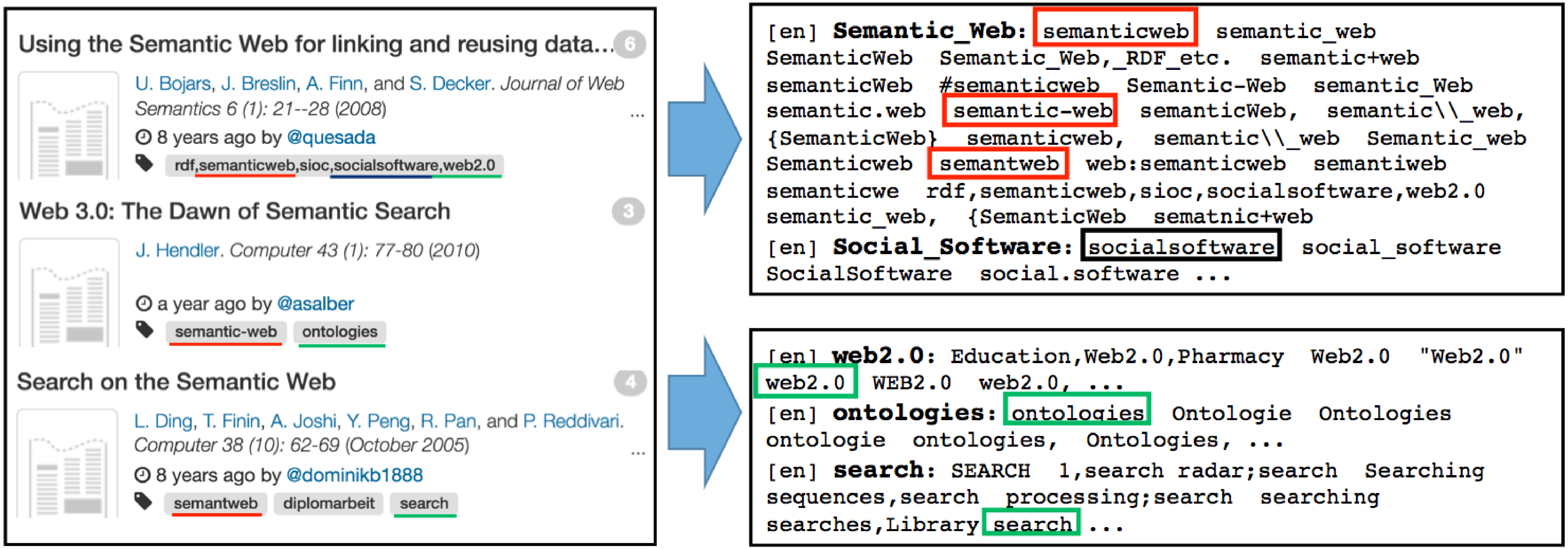


Image in Dong, H., Wang, W., & Coenen, F. (2017). Deriving Dynamic Knowledge from Academic Social Tagging Data: A Novel Research Direction. In iConference 2017 Proceedings (pp. 661-666). https://doi.org/10.9776/17313

- Positive data: tag concept pairs Ca, Cb
  - (i) satisfying *criteria* in the social tagging data, p(Ca|Cb) > TH
  - (ii) matched to a subsumption relation in any of the KBs.
- Negative data:
  - Reversed negative (if A->B is positive, then B->A is negative)
  - Random negative

Evaluation strategy

- Relation-level evaluation
  - Evaluate the classification model: results on testing data (held-out 20%)
  - **Outperformed all other baselines.**

- Ontology-level evaluation
  - Evaluate the generated hierarchies: using *Taxonomic* precision, recall, f-measure
  - Root concepts: Selected concepts under CS/IS categories in DBpedia and ACM.
  - Evaluate against sub-KBs. Averaging the *Taxonomic* precision, recall and calculate F-measure.
  - **Results not consistent, but our proposed approach has generally better/competitive results.**

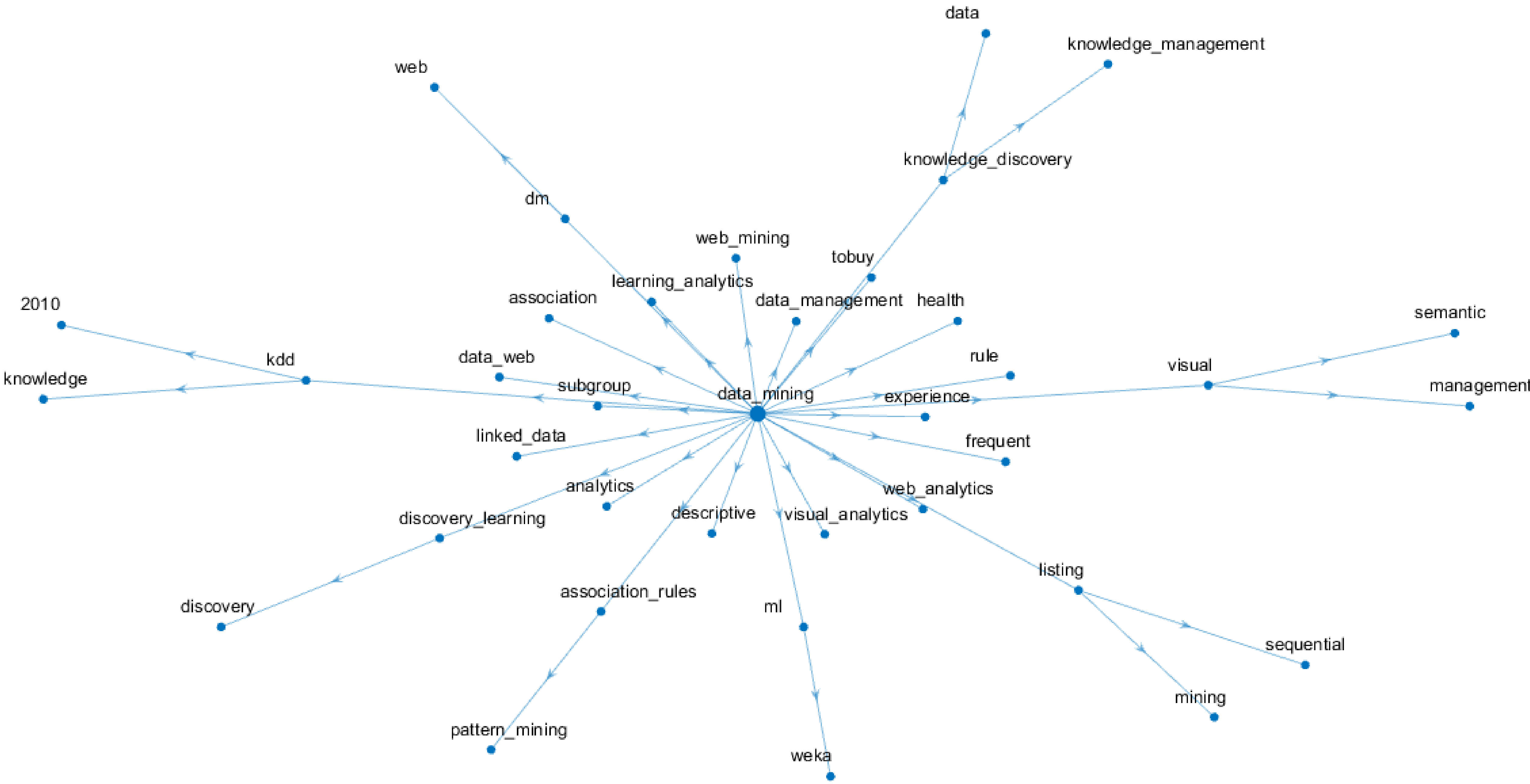
- Enrichment-based evaluation
  - Enriched 3846 relations to DBpedia and 1302 relations to ACM.
  - Selected 298 and manual evaluation by 7 experts, with our proposed approach, **41.18%** = 859/(298*7) are marked as subsumption, higher than 33.33% as random (3 categories to rate).

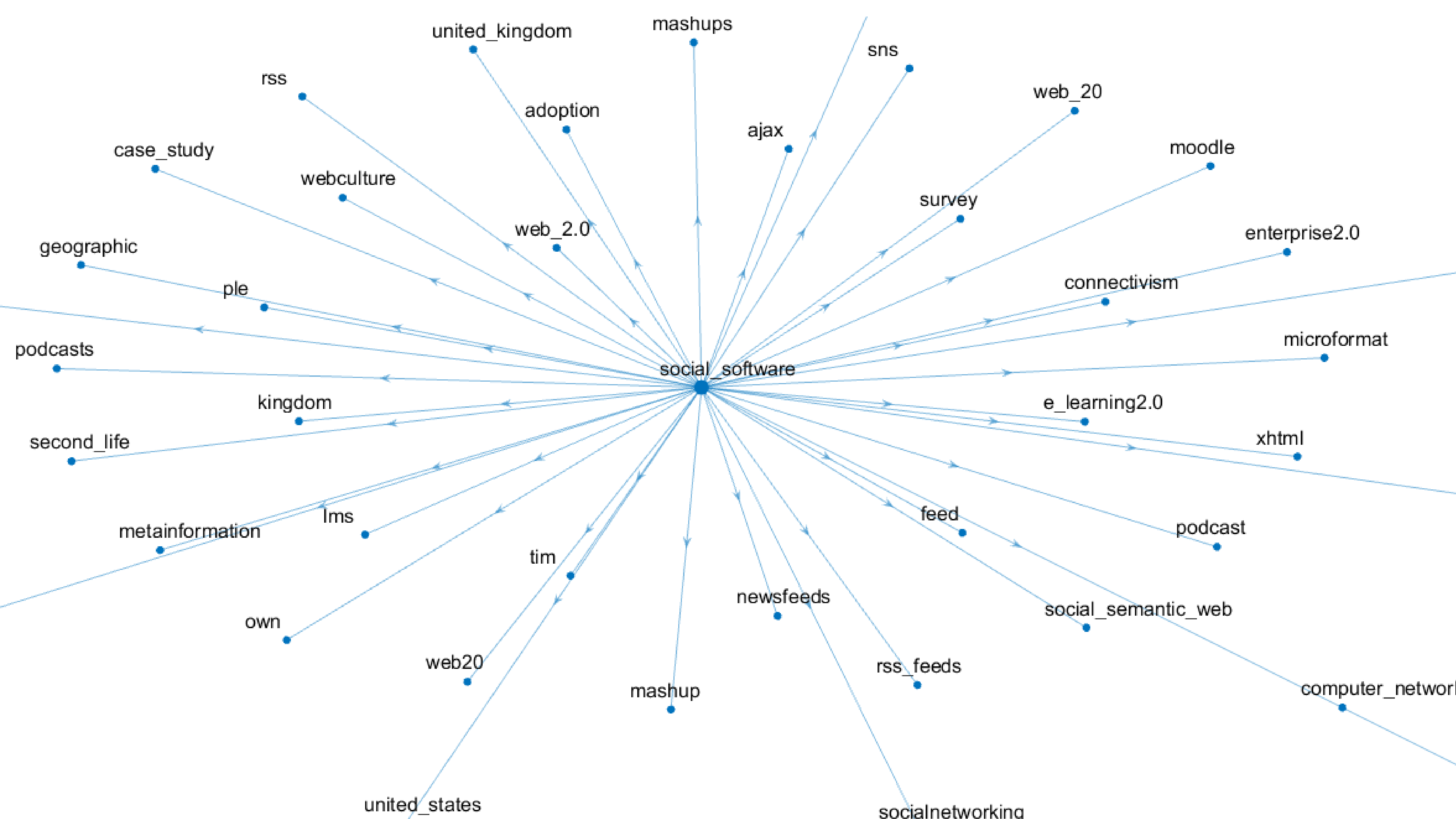
Results – Relation-level evaluation

Table 8: Classification Testing Results with Comparison among Feature Sets

		R	P	F1
Random setting		50.00%	34.16%	40.59%
$S_{all} = S_{sim} + S_{topic\text{-}dist} + S_{prob\text{-}asso}$ (Full features in our approach)	SVM RBF ($2^{10.5}$,$2^{4.5}$)	51.56%	52.95%	52.25%
	AdaBoost	50.15%	63.52%	**56.05%**
	LR	34.04%	65.00%	44.68%
	DT	45.02%	62.87%	52.46%
S_{CO} (Rêgo et. al [1])	SVM RBF (2^{10},2^{7})	36.96%	58.81%	**45.39%**
	AdaBoost	27.49%	61.07%	37.92%
	LR	19.64%	56.20%	29.10%
	DT	27.19%	58.95%	37.22%
S_{all+CO}	SVM RBF ($2^{9.5}$,2^{4})	49.25%	52.41%	50.78%
	AdaBoost	46.32%	65.25%	**54.18%**
	LR	36.56%	62.69%	46.18%
	DT	46.73%	57.35%	51.50%
S_{sim} (Wang et. al [33])	SVM RBF ($2^{10.5}$,2^{9})	46.02%	47.02%	**46.51%**
	AdaBoost	17.52%	59.59%	27.08%
	LR	15.01%	54.78%	23.56%
	DT	11.78%	66.10%	20.00%
$S_{topic\text{-}dist}$	SVM RBF (2^{10},2^{11})	40.28%	46.14%	**43.01%**
	AdaBoost	11.48%	59.07%	19.22%
	LR	10.27%	55.14%	17.32%
	DT	3.02%	47.62%	5.68%
$S_{prob\text{-}asso}$	SVM RBF (2^{12},$2^{8.5}$)	27.80%	60.53%	38.10%
	AdaBoost	44.51%	63.60%	52.37%
	LR	14.20%	68.12%	23.50%
	DT	53.07%	60.09%	**56.36%**

The values (2^a,2^b) after SVM RBF are the parameters c and γ tuned to optimise F_1 score.


data
knowledge_management
web
knowledge_discovery
dm
web_mining
tobuy
learning_analytics
association
data_management
health
2010
semantic
kdd
data_web
rule
visual
knowledge
subgroup
data_mining
experience
management
linked_data
frequent
analytics
web_analytics
descriptive
visual_analytics
discovery_learning
listing
association_rules
discovery
ml
sequential
mining
pattern_mining
weka


united_kingdom
mashups
sns
rss
web_20
adoption
ajax
case_study
moodle
webculture
survey
web_2.0
enterprise2.0
geographic
connectivism
ple
microformat
podcasts
social_software
kingdom
e_learning2.0
second_life
xhtml
feed
lms
podcast
metainformation
tim
newsfeeds
social_semantic_web
own
web20
rss_feeds
mashup
united_states
socialnetworking

Overview

- Relation learning: Automatic Taxonomy Generation from Social Tagging Data to Enrich Knowledge Bases
- **Tag Annotation: Sequence Modelling for Tag Annotation/Recommendation**

Research Tasks:

- Tag annotation: simulate human annotation process through a sequence model.
  - Reading a set of paragraphs and annotate them with tags/key words.

- Related tasks:
  - Tag recommendation - equivalent
  - Hashtag recommendation in microblog – related
  - Text summarisation – related but distinct (output is sequential)
  - Machine Translation – somehow related (output is sequential & different language)
  - Aspect-based sentiment classification? - maybe related (output is non-sequential but with probability/polarity)

Related work about attentions

- Neural Machine Translation by Jointly Learning to Align and Translate (Bahdanau, Cho & Benjio, ICLR 2015)

- Hierarchical Attention Networks for Document Classification (Yang *et al.*, NAACL-HLT 2016)

- Hashtag Recommendation with Topical Attention-Based LSTM (Li *et al.*, COLING 2016)

Attention Mechanism

- In NLP, firstly used in an encoder-decoder architecture for machine translation (Bahdanau, Cho & Benjio, 2015).

Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

Jane went to Africa last September, and enjoyed the culture and met many wonderful people; she came back raving about how wonderful her trip was, and is tempting me to go too.

Example in the online course [Sequence Models](), by Deeplearning.ai, Andrew Ng.

Attention Mechanism

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

$$e_{ij} = a(s_{i-1}, h_j)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

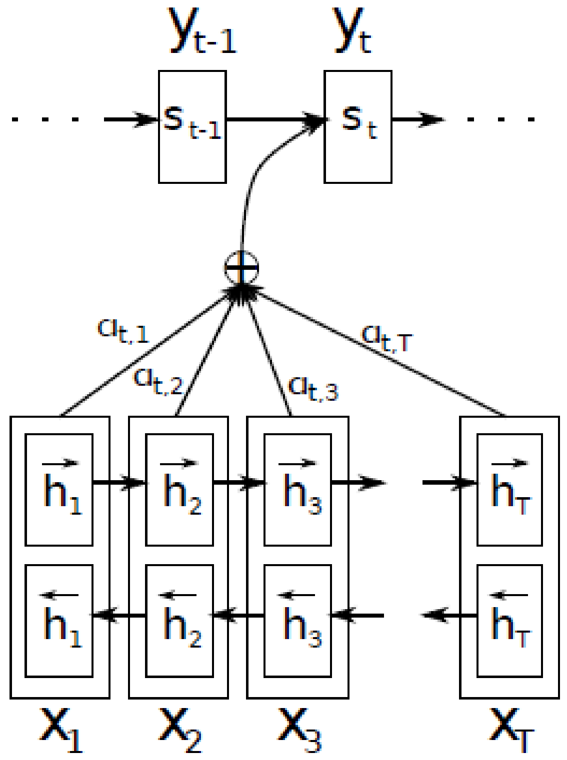


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Figure In Bahdanau, Cho & Bengio (2014).

Figure in (Yang et al., 2016)

Hierarchical Attention

From sentence to document

$$u_i = \tanh(W_s h_i + b_s),$$
$$\alpha_i = \frac{\exp(u_i^\top u_s)}{\sum_i \exp(u_i^\top u_s)},$$
$$v = \sum_i \alpha_i h_i,$$

From word to sentence

$$u_{it} = \tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$
$$s_i = \sum_t \alpha_{it} h_{it}.$$

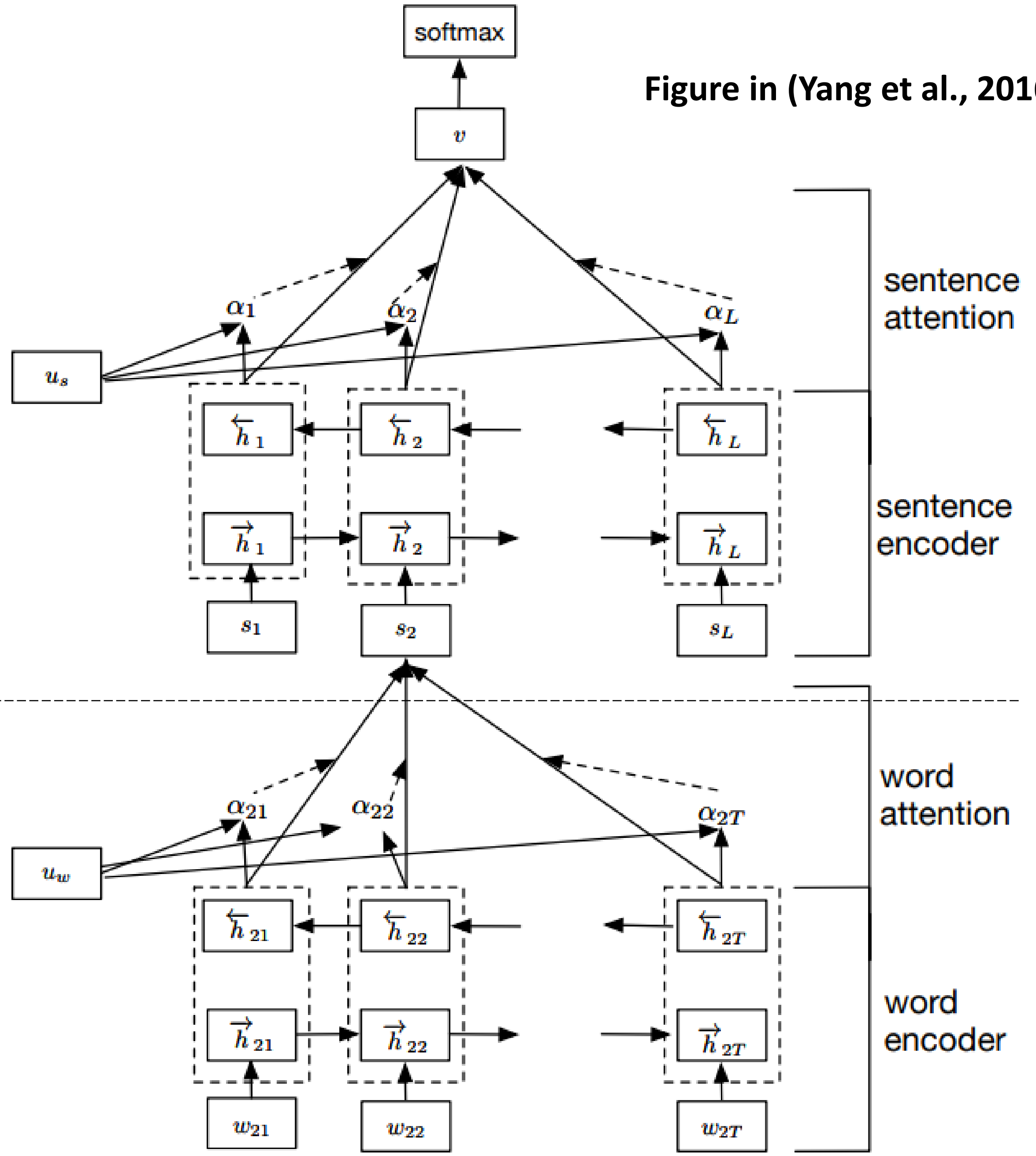


Figure 2: Hierarchical Attention Network.

Tables in (Yang et al., 2016)

Hierarchical Attention

- Measured with sentiment estimation & topic classification tasks

Data set	classes	documents
Yelp 2013	5	335,018
Yelp 2014	5	1,125,457
Yelp 2015	5	1,569,264
IMDB review	10	348,415
Yahoo Answer	10	1,450,000
Amazon review	5	3,650,000

	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

Table 2: Document Classification, in percentage

Figure in (Yang et al., 2016)

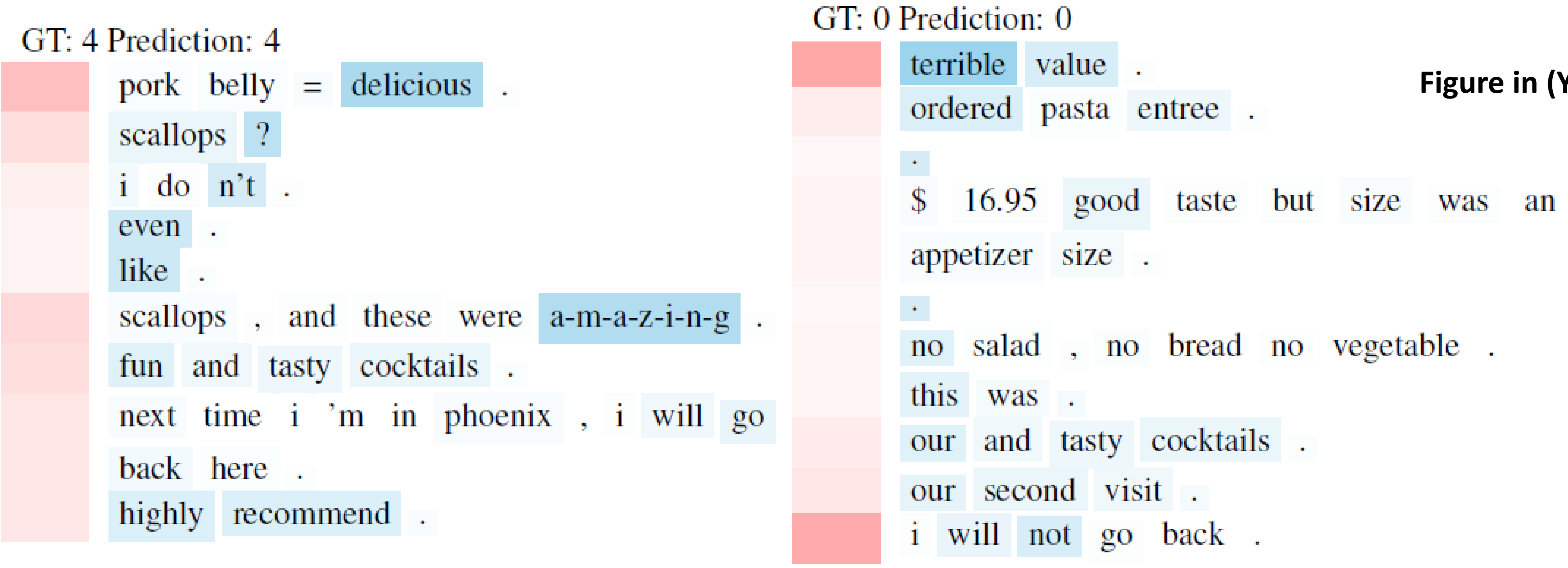


Figure 5: Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.

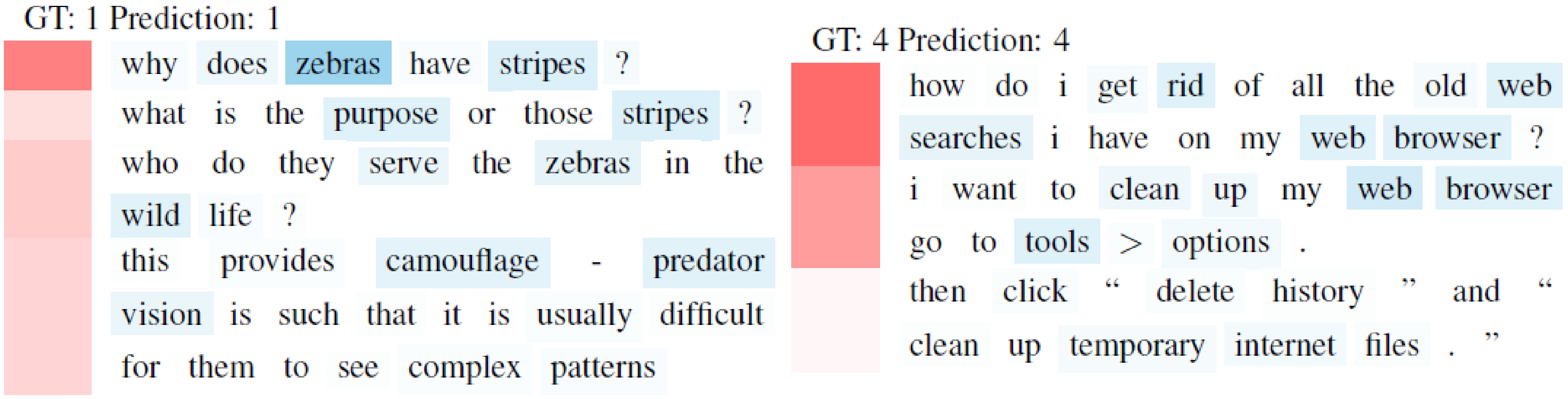


Figure 6: Documents from Yahoo Answers. Label 1 denotes Science and Mathematics and label 4 denotes Computers and Internet.

Topical Attention: Scenario and hypothesis

The topic information matters when generating hashtags.

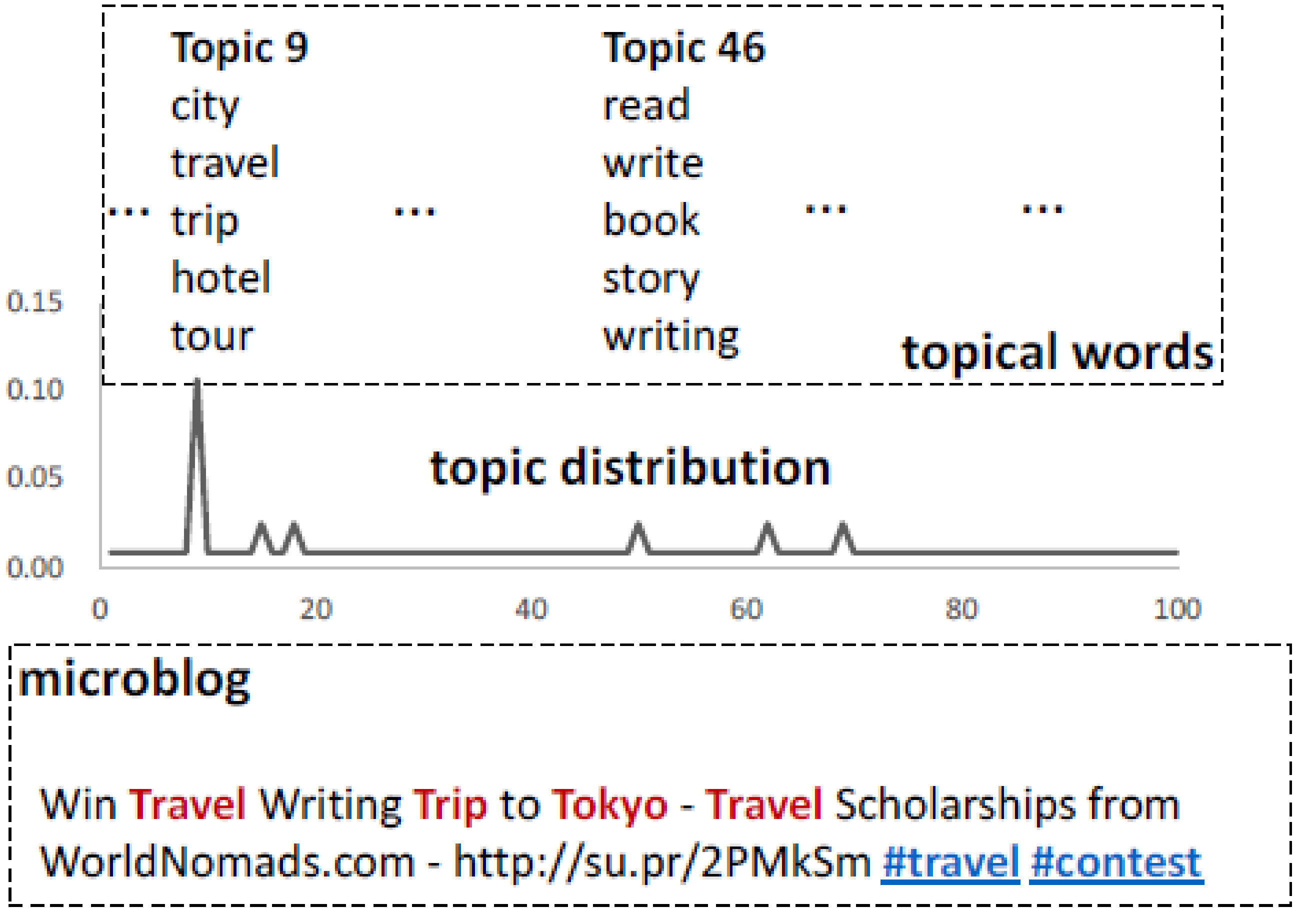


Figure in (Li et al., 2016)

Topical Attention

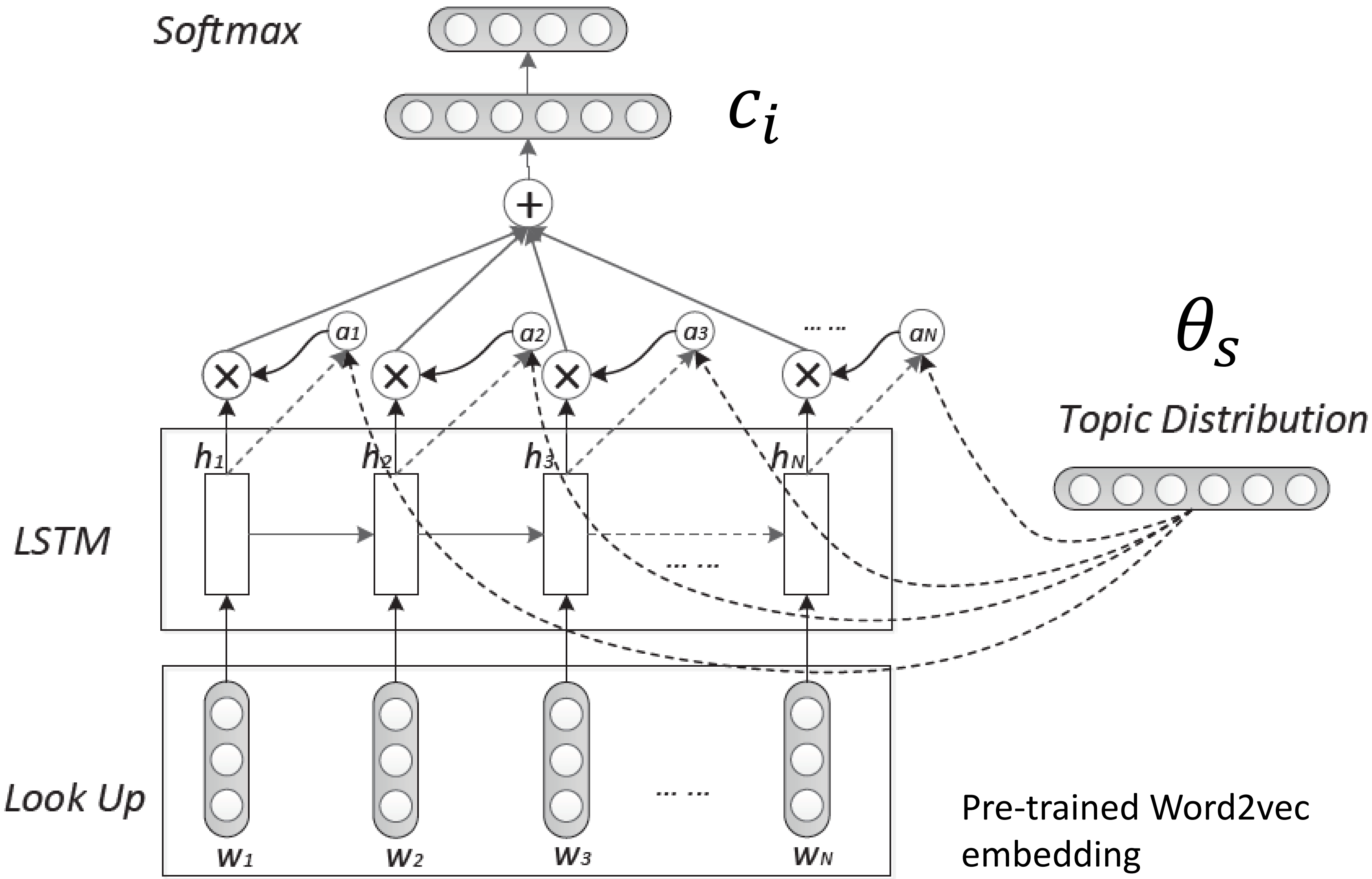
Figure in (Li et al., 2016)

- Topical Attention in a many-to-one RNN.

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j .$$

$$e_{ij} = a(s_{i-1}, \theta_s)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$



Dataset used

- Twitter dataset
- 185,291,742 tweets from Oct 2009 to Dec 2009, among them 16,744,189 tweets have hashtags annotated by users.

- Randomly selected 500,000 for training, 50,000 for development, 50,000 for testing.

# Tweets	# Hashtags	Vocabulary Size	Nt(avg)
600,000	27,720	337,245	1.308

Table 1: Statistics of the dataset, Nt(avg) is the average number of hashtags in the dataset.

Table in (Li et al., 2016)

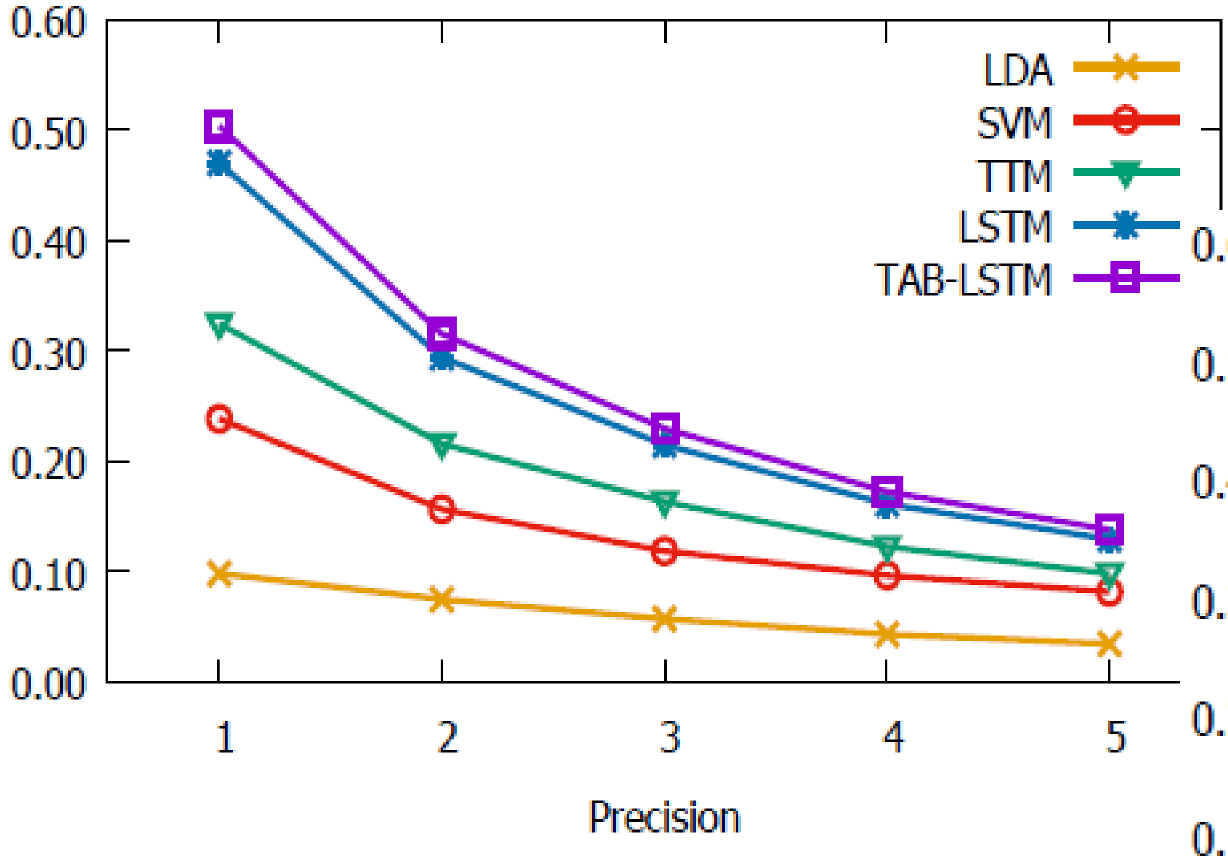
Results

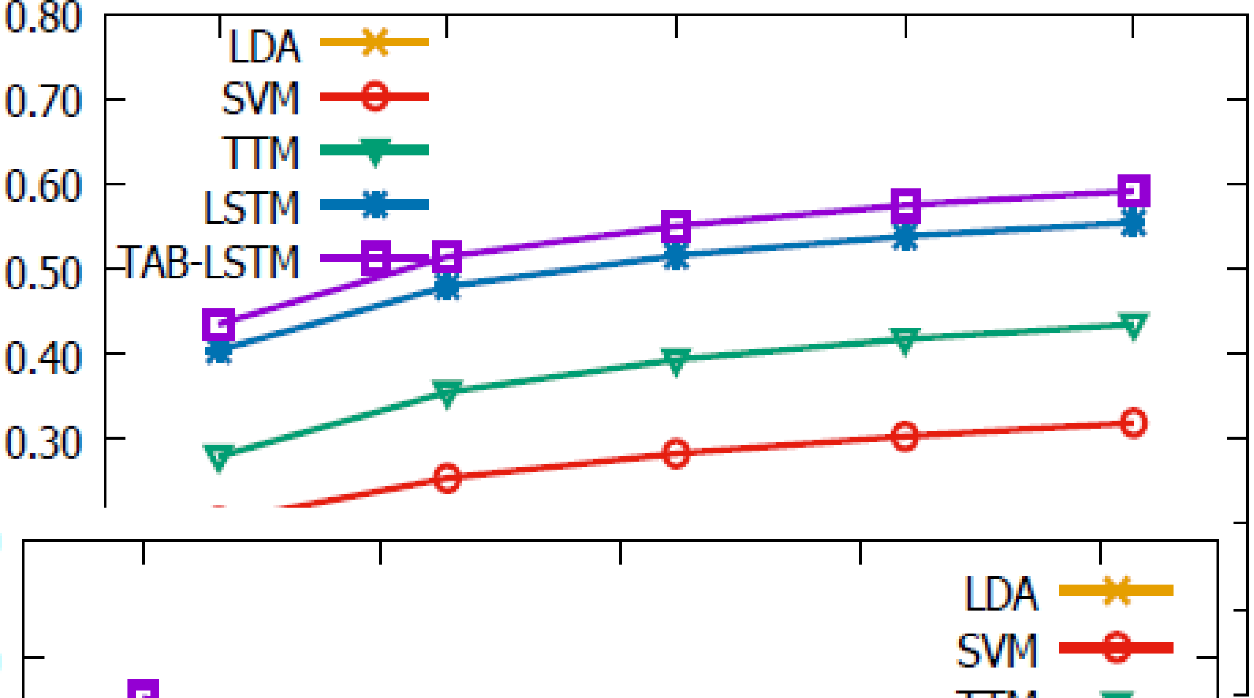
Table in (Li et al., 2016)

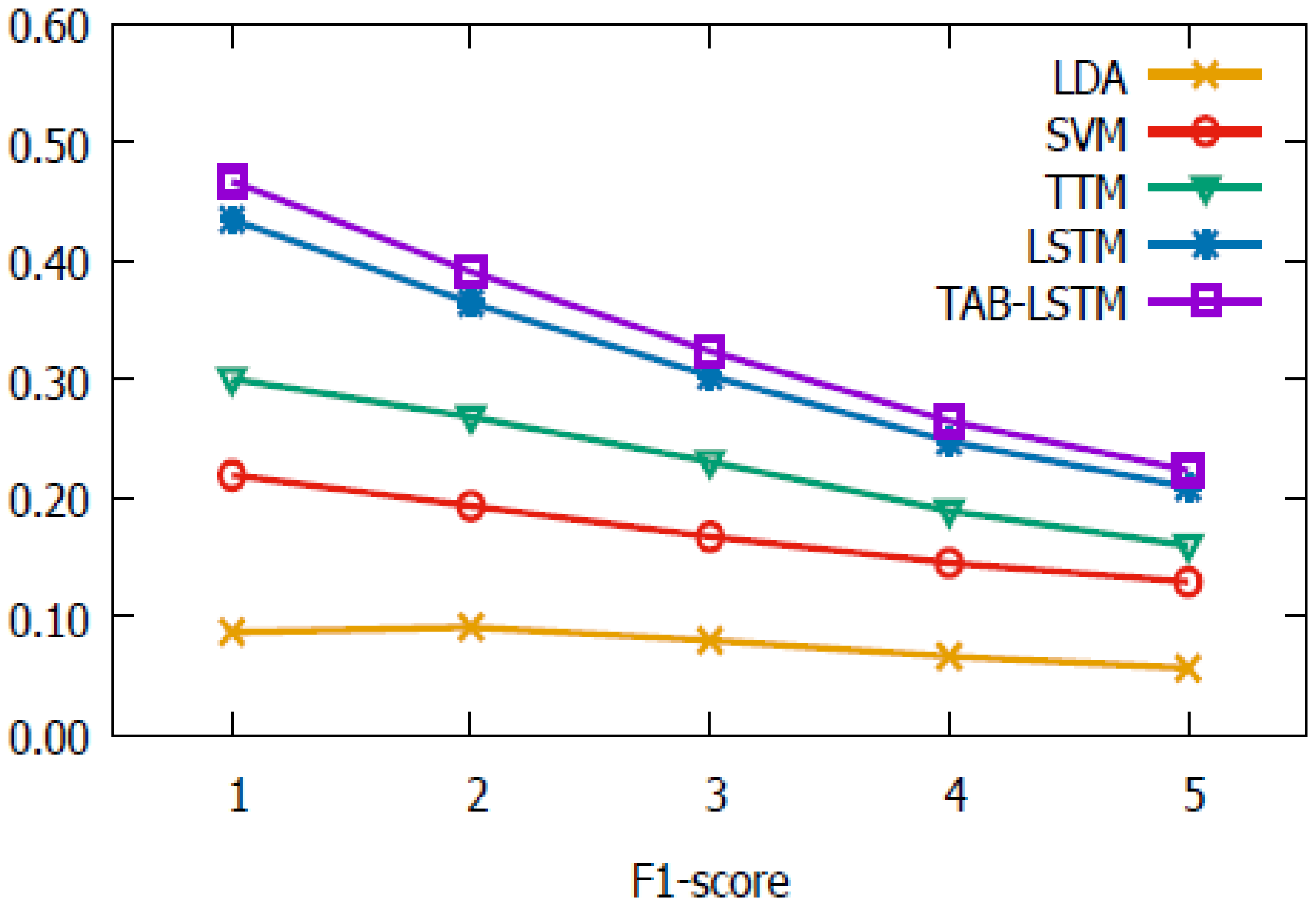
Methods	Precision	Recall	F1-score
LDA	0.098	0.078	0.087
SVM	0.238	0.203	0.219
TTM	0.324	0.280	0.300
LSTM	0.470	0.404	0.434
AVG-LSTM	0.472	0.405	0.436
VAB-LSTM	0.489	0.419	0.452
TAB-LSTM	**0.503**	**0.435**	**0.467**

Table 2: Evaluation results of different methods for hashtag recommendation. The dimension of word embeddings is set to be 300 for all methods. All improvements obtained by TAB-LSTM over other methods are statistically significant within a 0.99 confidence interval using the t-test.

Results (2)







Figures in (Li et al., 2016)

Visualisation of attention

Figure in (Li et al., 2016)

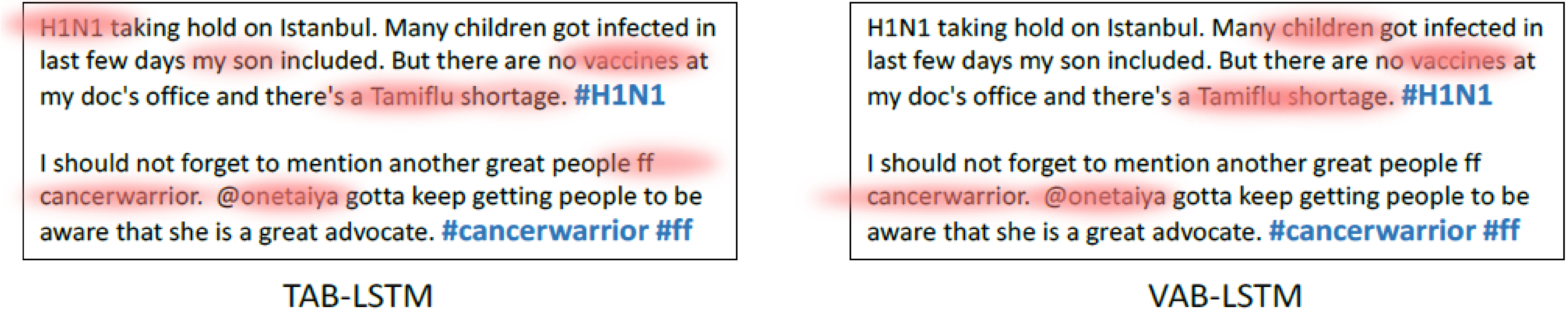


Figure 4: Attention heat maps for two example microblog posts.

Probably visualized using α_{ij} in the equation $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$

Back to my research

- Design a new attention mechanism suitable for social tag annotation.
- Understand the processing of tagging, taking temporal factors into consideration.

Key References

- T. V. Wal, “Folksonomy,” http://vanderwal.net/folksonomy.html, 2007.
- H. Dong, W. Wang and H.-N. Liang, "Learning Structured Knowledge from Social Tagging Data: A Critical Review of Methods and Techniques," *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, Chengdu, 2015, pp. 307-314.
- H. Dong, W. Wang, F. Coenen. Deriving Dynamic Knowledge from Academic Social Tagging Data: A Novel Research Direction, *iConference 2017*, Wuhan, P.R. China, 2017.3.22-3.25.
- A. Garcia-Silva, O. Corcho, H. Alani, and A. Gómez-Pérez, “Review of the state of the art: discovering and associating semantics to tags in folksonomies,” *The Knowledge Engineering Review*, vol. 27, no. 1, p. 5785, 2012.
- D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- A. S. C. Rego, L. B. Marinho, and C. E. S. Pires, “A supervised learning approach to detect subsumption relations between tags in folksonomies,” in *Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC ’15)*. ACM, 2015, pp. 409–415.
- J. Chen, S. Feng, and J. Liu, “Topic sense induction from social tags based on non-negative matrix factorization,” *Information Sciences, vol. 280, pp. 16-25, 2014.*
- P. Andrews, and J. Pane, “Sense induction in folksonomies: a review,” *Artificial Intelligence Review*, vol. 40, no. 2, pp. 147-174, 2013.
- M. Zhou, S. Bao, X. Wu, and Y. Yu, "An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations," Berlin, Heidelberg, 2007, pp. 680-693: Springer Berlin Heidelberg.
- D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
- Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480-1489.
- Y. Li, T. Liu, J. Jiang, and L. Zhang, "Hashtag Recommendation with Topical Attention-Based LSTM," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3019-3029.