Karlsruhe Institute of Technology

Unsupervised Multi-Topic Labeling for Spoken Utterances

Sebastian Weigelt, Jan Keim, Tobias Hey, and Walter F. Tichy

KIT – Department of Informatics – Institute for Program Structures and Data Organization (IPD).

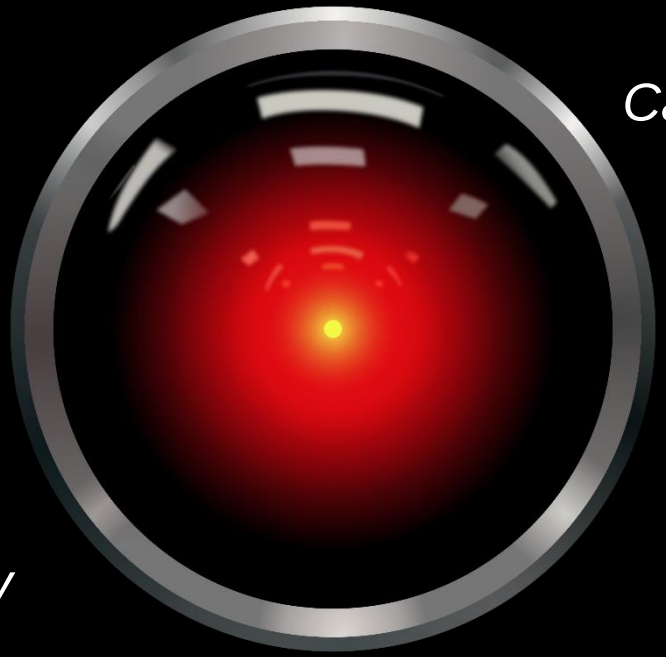
Intelligent Systems seem to be quite smart today!

How far away is the sun?

When is my next meeting?

Is my daughter at home?

Set a timer for five minutes



Call my brother at work

Remind me to call mom

How many dollars is 45 euro

Who is near me?

Google the war of 1812

Text Brian I'm on my way

Give me directions home

Find Disney movies

Play iTunes Radio

Should I bring an umbrella?

What's trending on twitter?

But do they really understand?

NO!

(they have no clue what you’re talking about!)

HEY ROBOT TAKE

UHM

~~THE APPLE~~

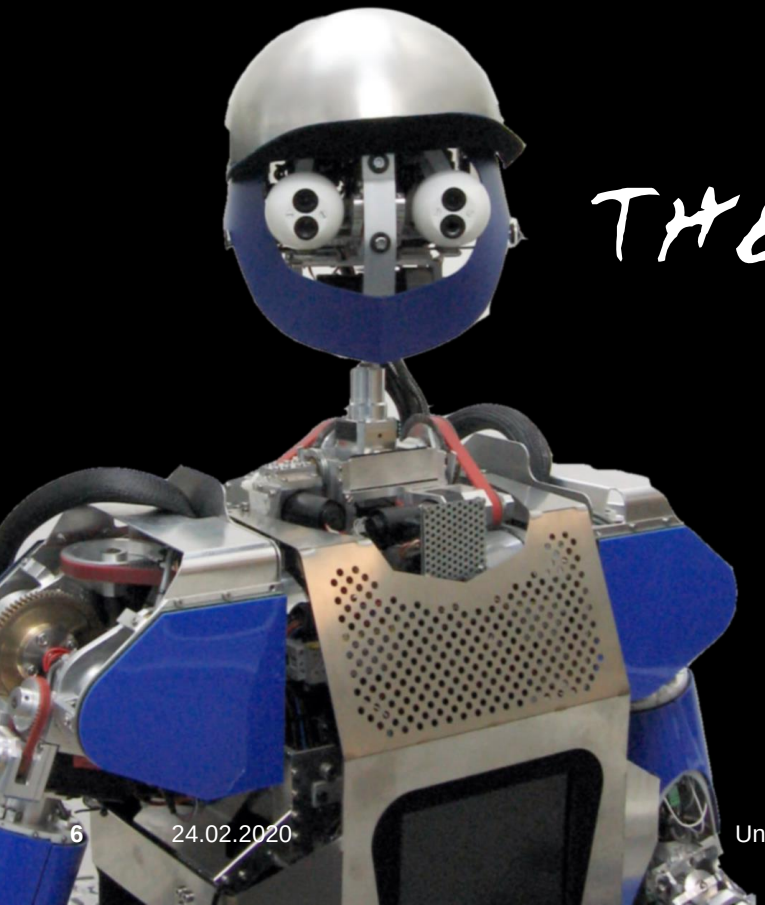
ERR

THE ORANGE FROM

THE FRIDGE

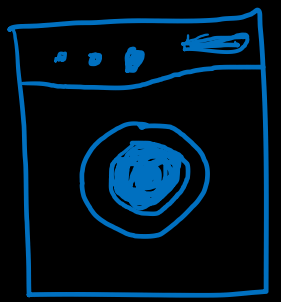
AND CLOSE THE DISHWASHER

AFTERWARDS



robotics

fruits



home appliances

Topic Extraction?

Wasn’t that solved in the 90s?

Yes but…

…for text (documents) only
…respectively for document sets!

HEY ROBOT TAKE

UHM

~~THE APPLE~~

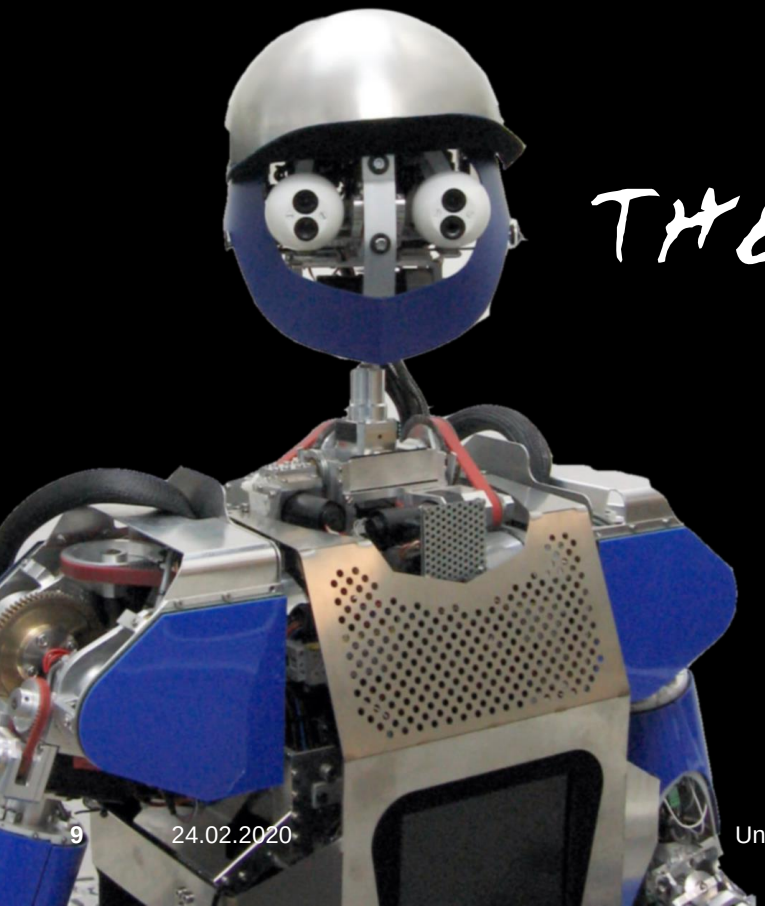
ERR

THE ORANGE FROM

THE FRIDGE

AND CLOSE THE DISHWASHER

AFTERWARDS



UNGRAMMATICAL

SHORT

ERRONEOUS

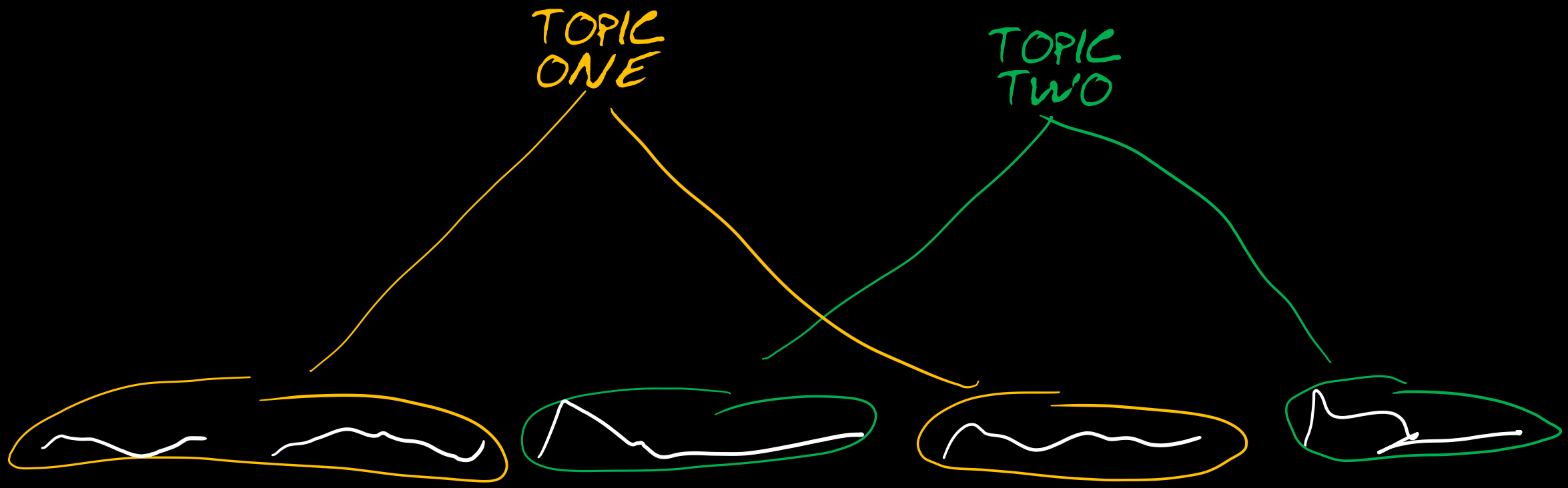
Related Work

What‘s the state of the art?

Related Work – Topic Extraction on Documents

Related Work – Topic Clustering on Speech

Cerisara[5], Hazen et al. [6], Siu et al. [7], …



Approach

What‘s the concept?

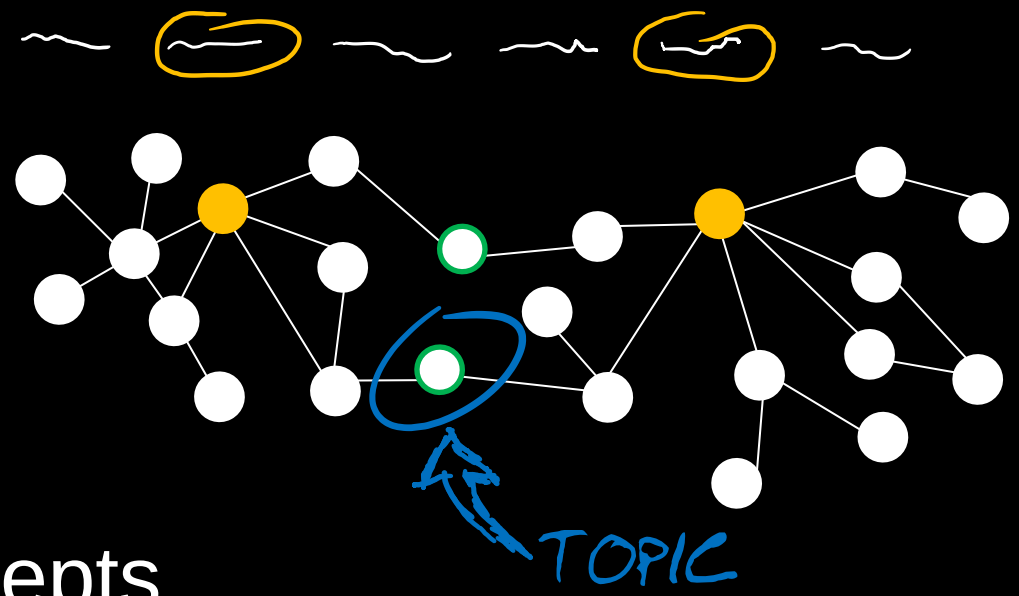
Approach – Overview

Unsupervised Multi-Topic Labeling for Spoken Utterances

(1) Disambiguate Terms

(2) Build Semantic Graphs

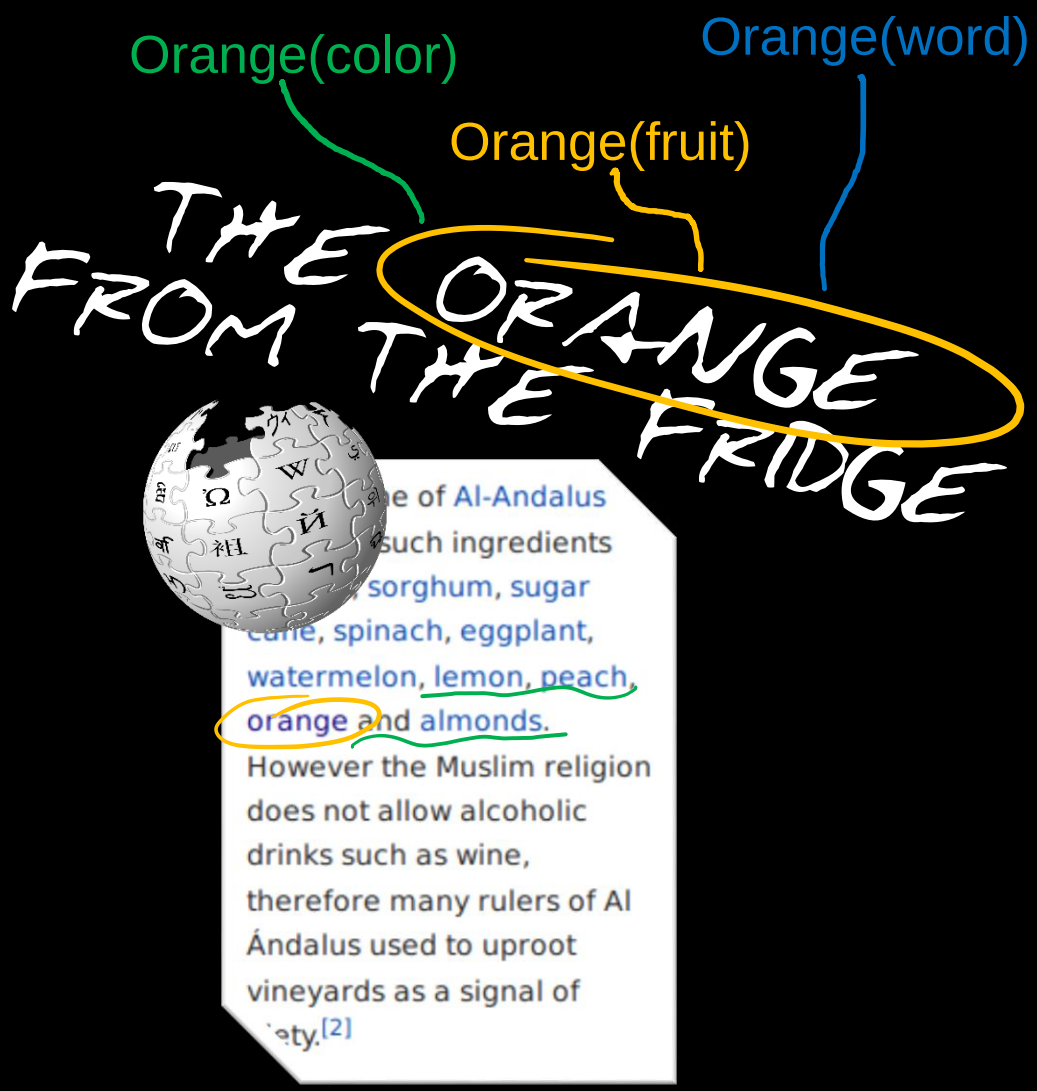
(3) Determine Most Central Concepts



Word Sense Disambiguation

What‘s that thing?

Approach – Disambiguation



Use Wikipedia as labeled corpus

- links are disambiguated senses
- extract all links from all articles

Train a classifier

- features: surrounding words, PoS tags, …
- classifier: naïve-bayes

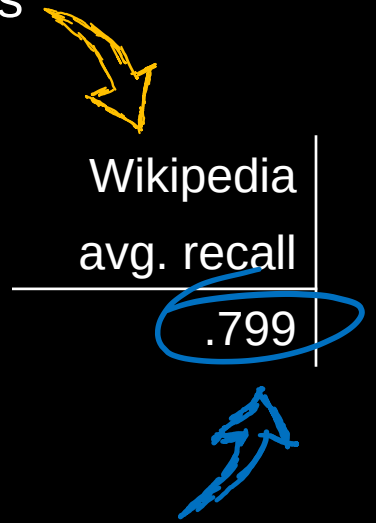
Proposed by Mihalcea et al.[8,9]

Approach – Disambiguation (Evaluation)

customized ten-fold cross

- ten runs
- 10,000 random instances each

= 100,000 instances



Mihalcea and Csomai – recall: **.831**

- 85 random articles
- 7,286 instances

Approach – Disambiguation (Evaluation)

customized ten-fold cross
- ten runs
- 10,000 random instances each

= 100,000 instances

Wikipedia	Speech Corpus		
avg. recall	precision	recall	F_1
.799	.894	.876	.885

manual transcriptions
- 168 recordings
- 1,060 nouns
- instances must be predicted

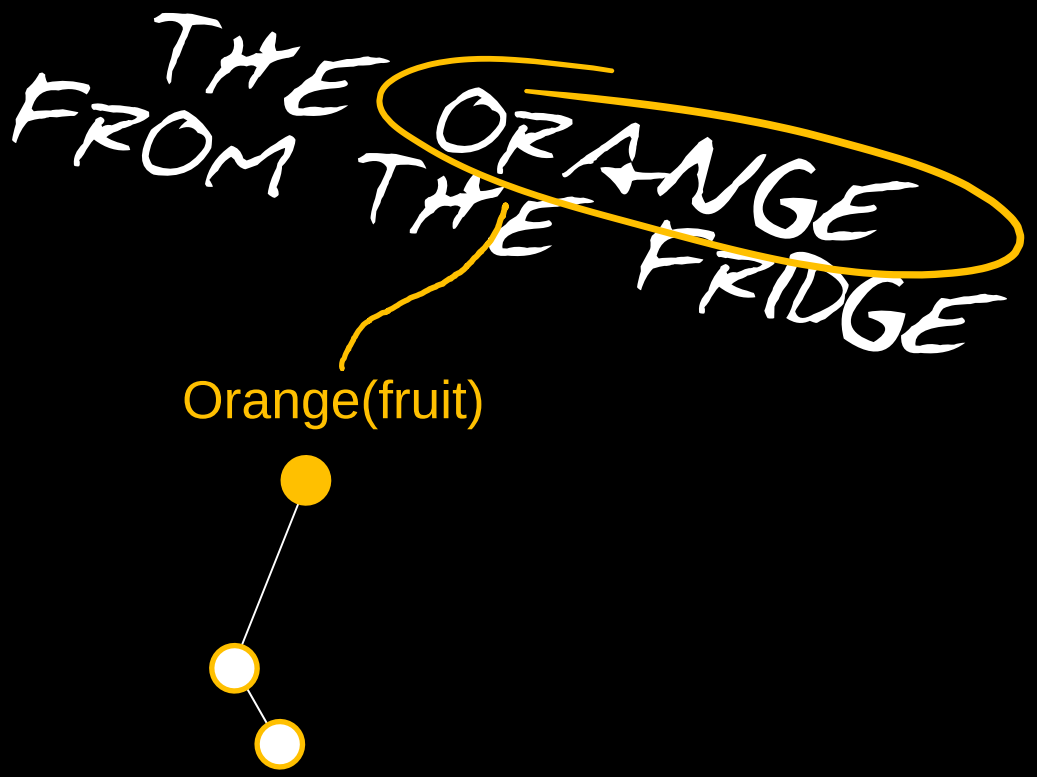
Mihalcea and Csomai – recall: **.831**
- 85 random articles
- 7,286 instances

encouraging result
- even better
- different domain

Semantic Graphs

What‘s the context?

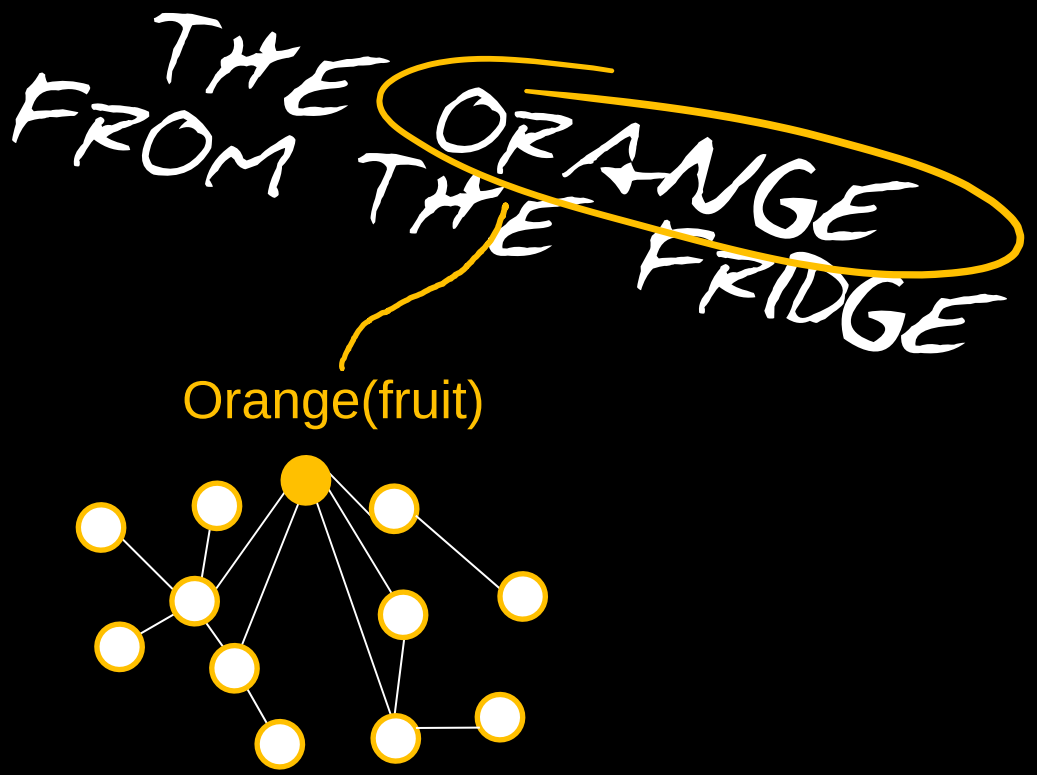
Approach – Semantic Graphs



Extract sub-graphs from DBpedia

- concepts are nodes
- relations are vertices
- depth of two

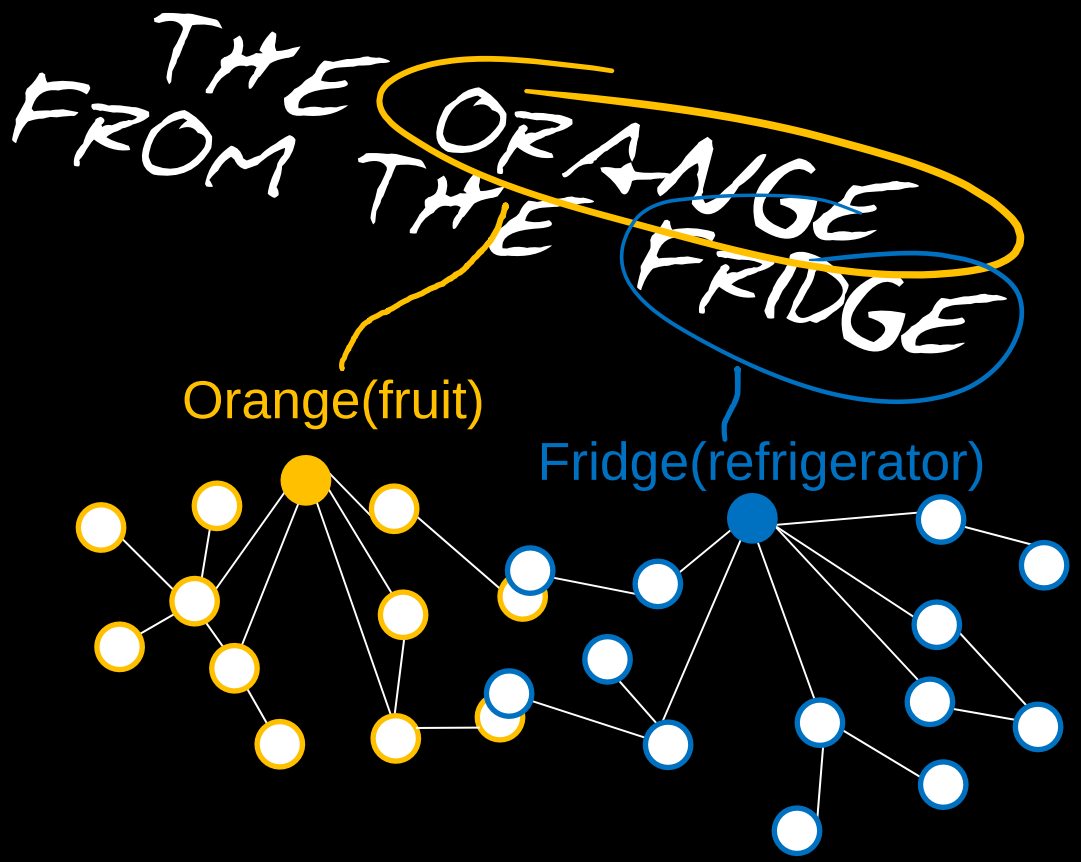
Approach – Semantic Graphs



Extract sub-graphs from DBpedia

- concepts are nodes
- relations are vertices
- depth of two

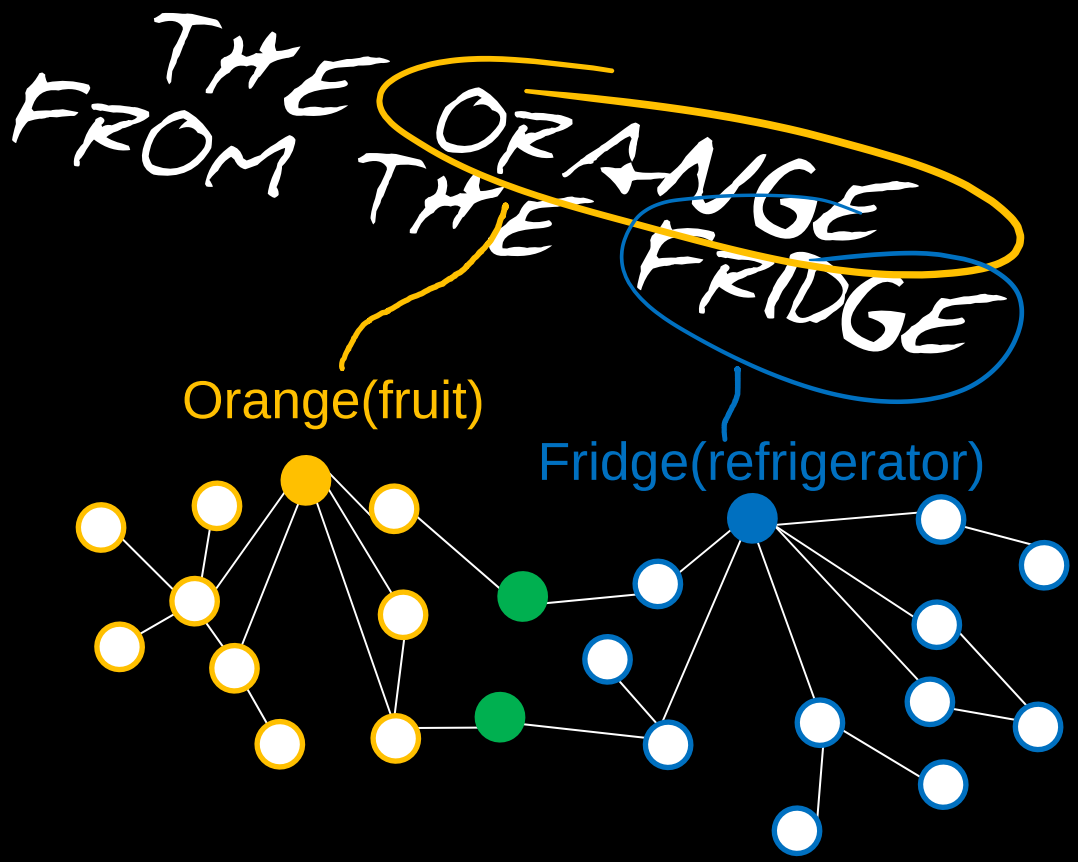
Approach – Semantic Graphs



Extract sub-graphs from DBpedia

- concepts are nodes
- relations are vertices
- depth of two

Approach – Semantic Graphs



Extract sub-graphs from DBpedia

- concepts are nodes
- relations are vertices
- depth of two

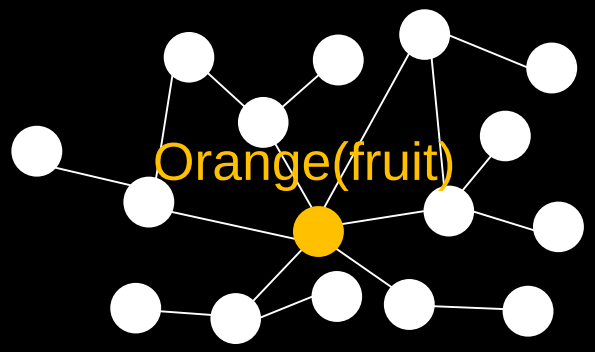
Join the semantic graphs

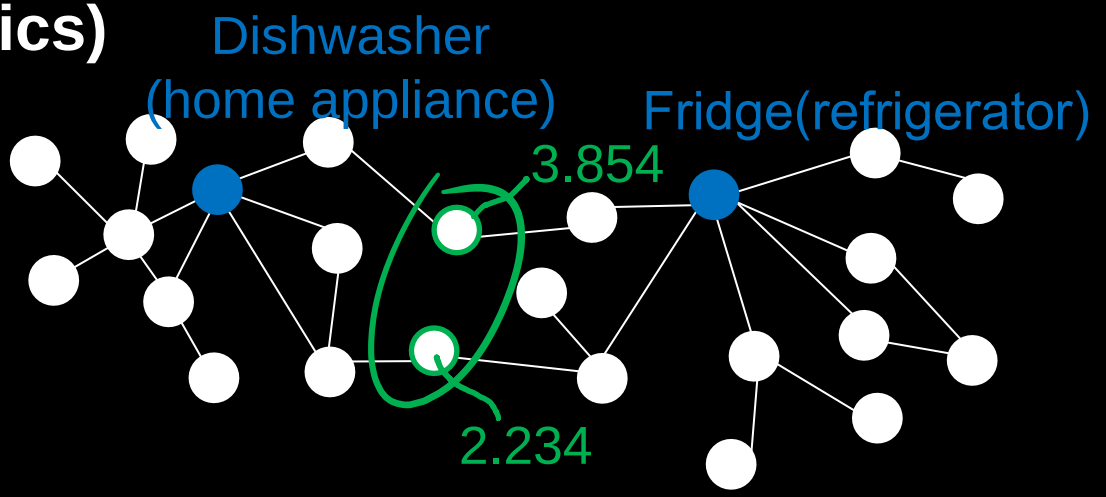
- at shared nodes
- implicit topic modeling

Central Terms

What‘s the topic?

Approach – Central Terms (Topics)





Determine central nodes

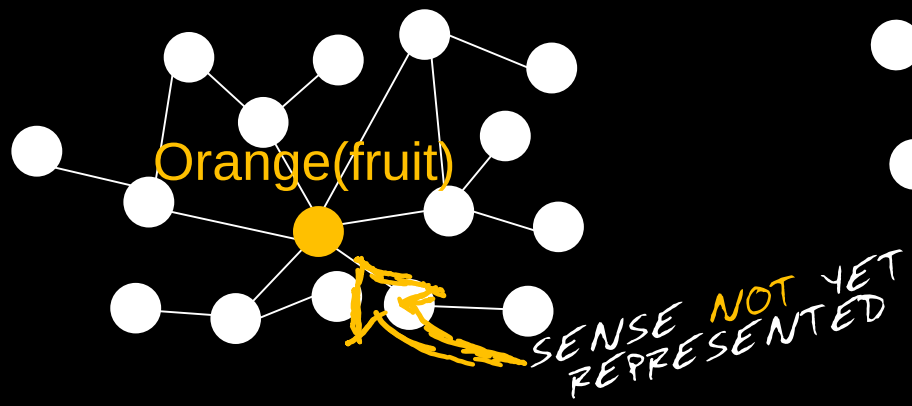
- algorithms require connected graphs
- use node connectivity and the biased PageRank algorithm instead[4,10]

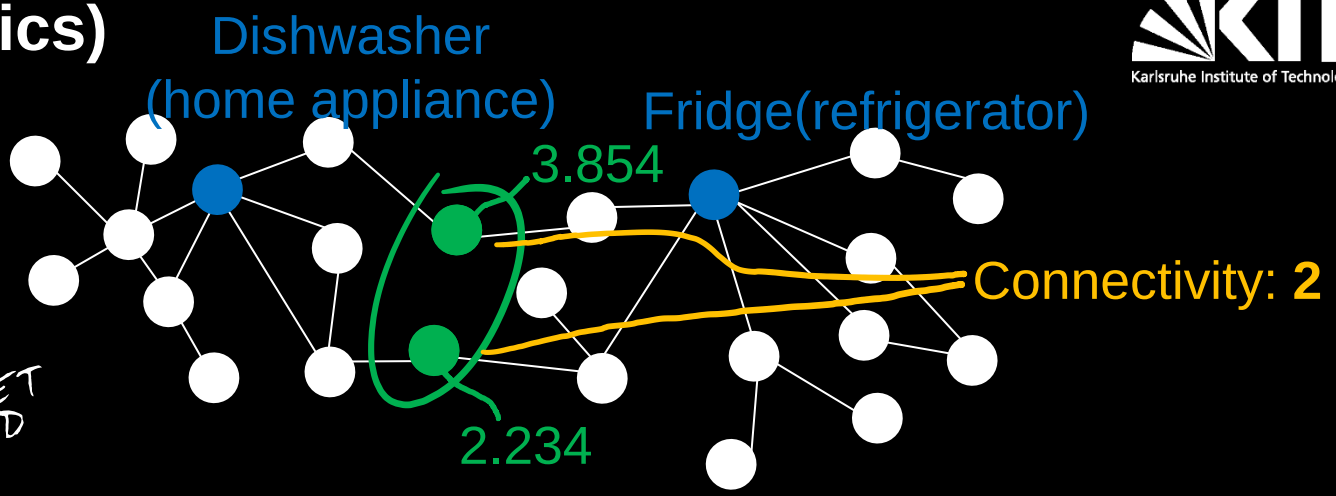
$$S(V_i) = (1-d) * B(V_i) + d * \sum_{j \in I(V_i)} \frac{S(V_j)}{|O(V_j)|}$$

$$B(V_i) = \begin{cases} 0 & , V_i \notin \text{InitNodes} \\ \frac{1}{|\text{InitNodes}|} & , V_i \in \text{InitNodes} \end{cases}$$

FOR ALL THE MATH NERDS OUT THERE

Approach – Central Terms (Topics)





Top Connectivity Strategy

- select twice as many topics than (distinct) senses
- select nodes with highest connectivity
- draw: use PageRank value

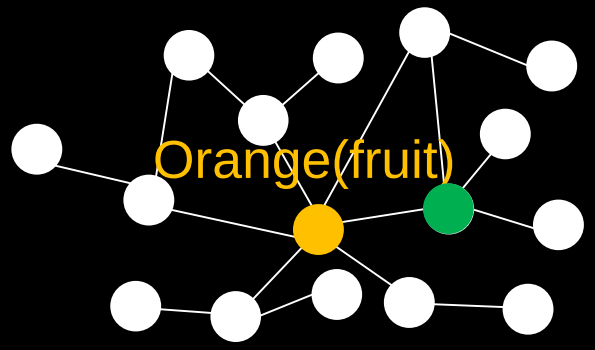
Topic one:

Home appliances

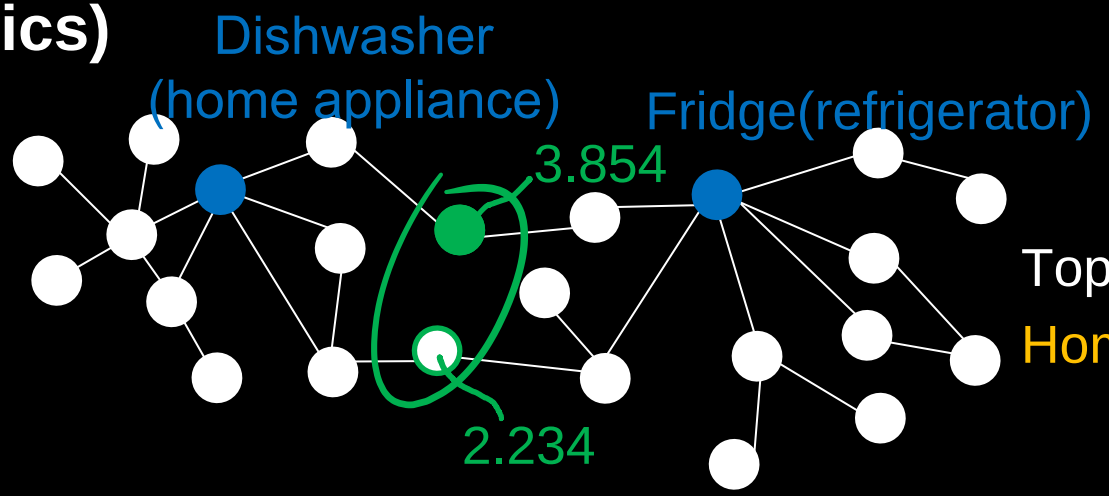
Topic two:

Home

Approach – Central Terms (Topics)



Topic two:
Fruit



Topic one:
Home appliances

Top Connectivity Strategy

- select twice as many topics than (distinct) senses
- select nodes with highest connectivity
- draw: use PageRank value

Max Coverage Strategy

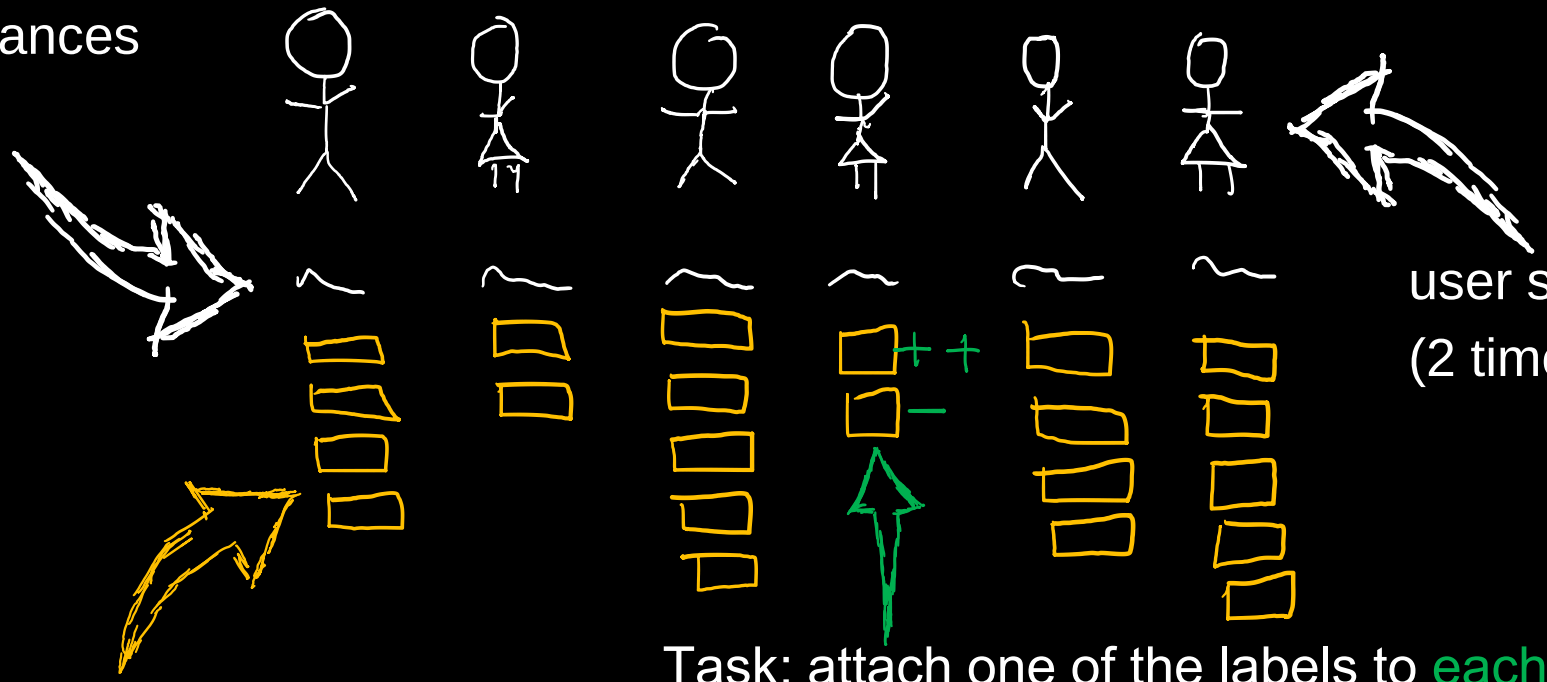
- select top connected nodes for all sub-graphs
- fill remaining places with overall top connected nodes

Evaluation

Will it work?

Evaluation – Set-Up

22 sample utterances
(4 domains)



user study with 6 subjects
(2 times 3 each)

extracted topics
(4 to 10, ordered)

Task: attach one of the labels to each topic

- good fit + +
- too broad +
- inconvenient –
- unrelated – –

Evaluation – Distribution of Labels

k	good fit		too broad		inconv.		unrelated	
	max	top	max	top	max	top	max	top

Evaluation – Distribution of Labels

k	good fit		too broad		inconv.		unrelated	
	max	top	max	top	max	top	max	top
1	.530	.530	.242	.242	.045	.045	.182	.182

more than half of the top ranked labels are a good choice…

…and three-quarters are okay at least

Evaluation – Distribution of Labels

k	good fit max	good fit top	too broad max	too broad top	inconv. max	inconv. top	unrelated max	unrelated top
1	.530	.530	.242	.242	.045	.045	.182	.182
2	.447	.424	.167	.182	.106	.106	.280	.288
3	.449	.444	.141	.152	.157	.146	.253	.258
4	.432	.420	.144	.140	.155	.167	.269	.273
5	.381	.369	.145	.136	.176	.182	.298	.312
all	.368	.340	.138	.132	.179	.200	.315	.329

BEST = BLUE

the more labels, the more opportunities to fail

the max strategy outperforms top (almost always)

Evaluation – Distribution of Labels

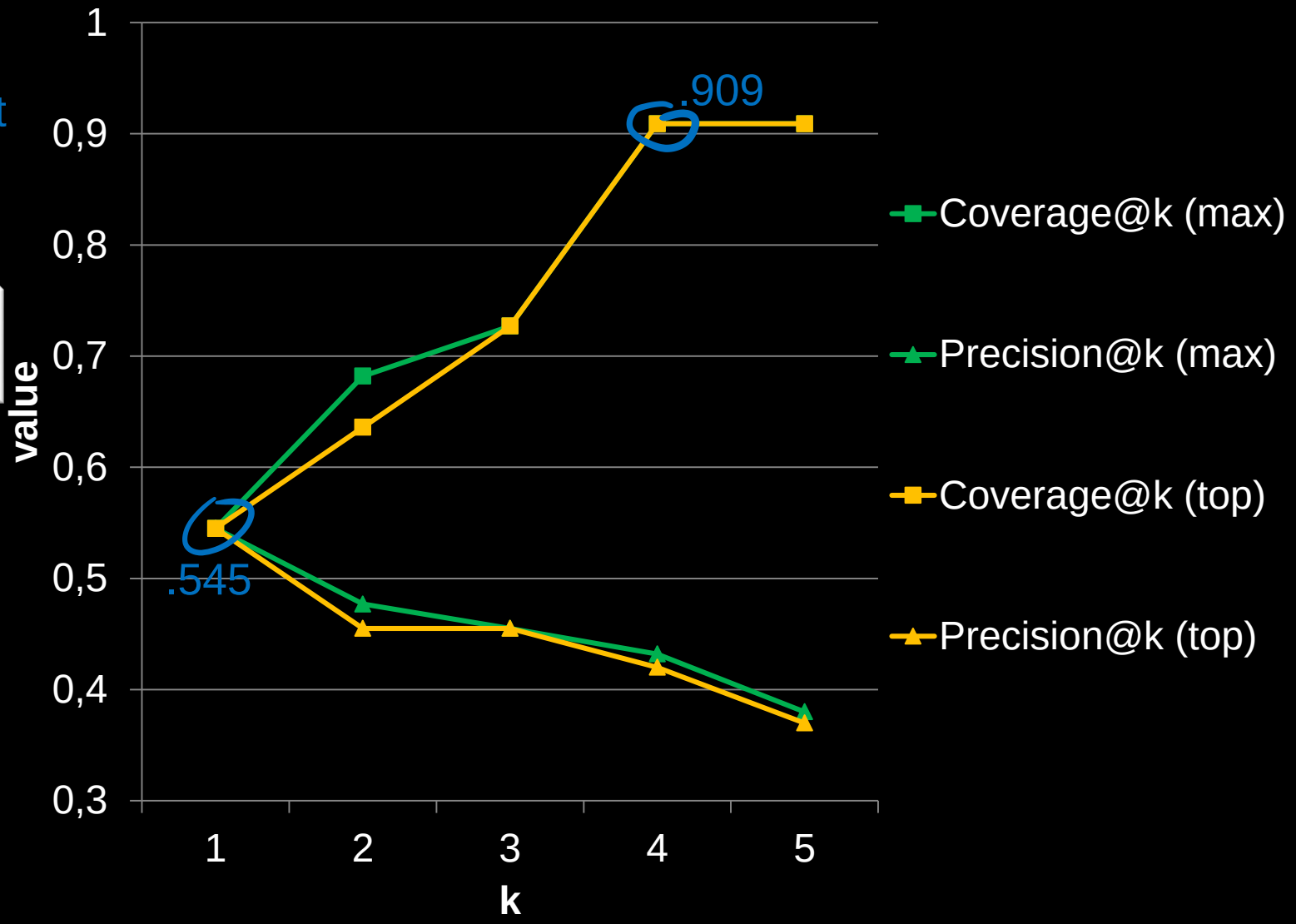
What about majority decisions?

- The ratio of utterances having at least one fitting label at rank *k* (min two of three subjects)

$$C@k = \frac{\#\text{u. w/ at least 1 } Hit \text{ at rank} \leq k}{\#\text{utterances}}$$

- The ratio of fitting labels among the *k* labels

$$P@k = \frac{\#Hits \text{ with rank} \leq k}{k}$$

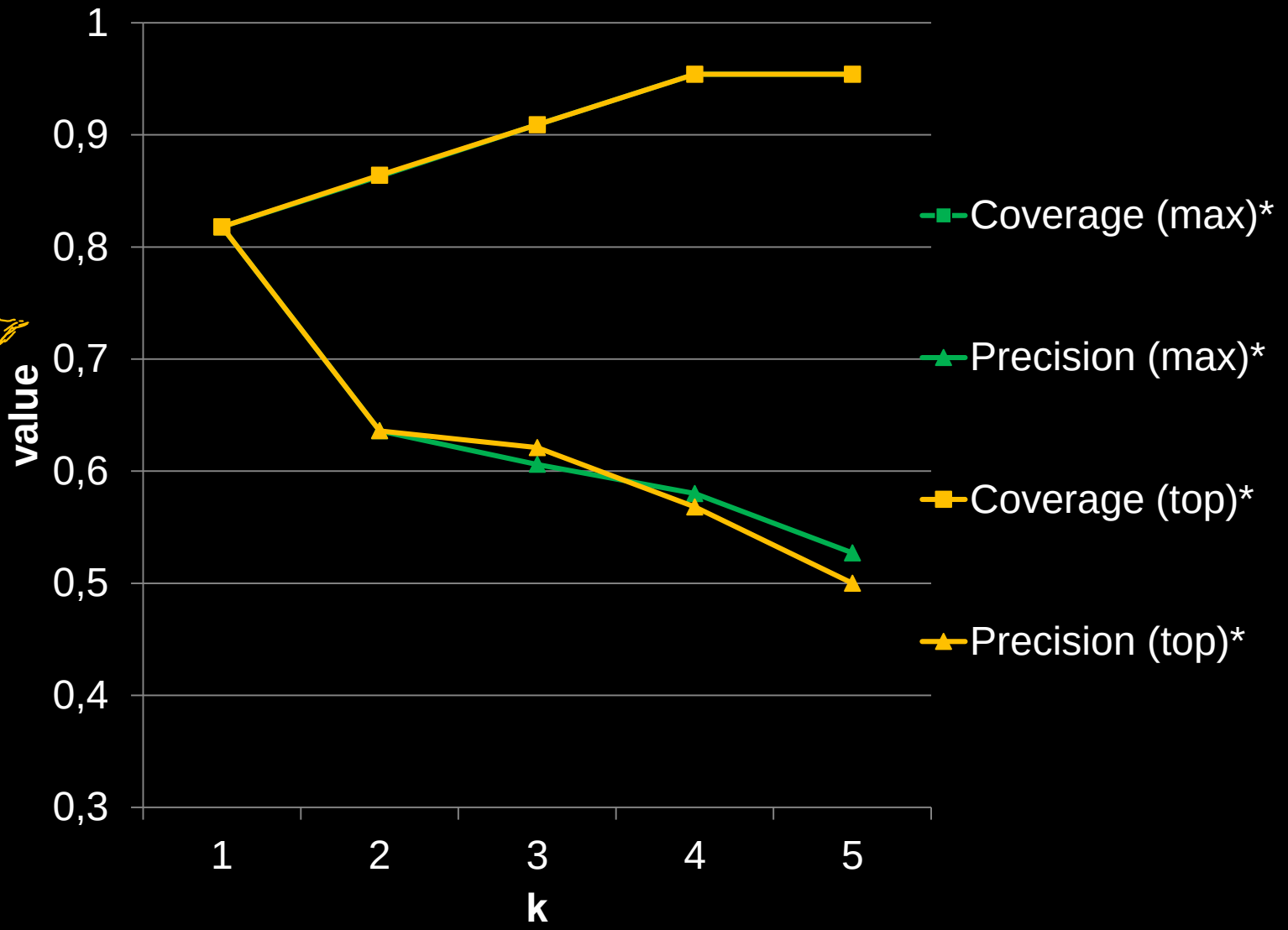


Evaluation – Distribution of Labels

Good fit or broader?

- pretty much the same
- above .8 for k = 1

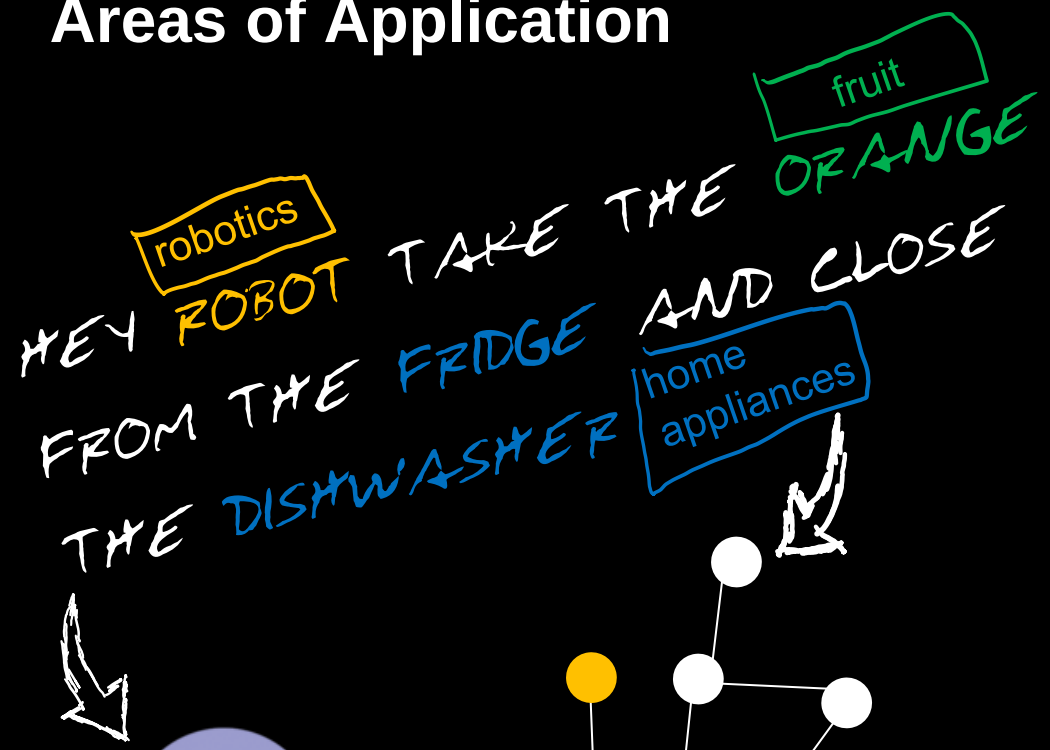
fortunately too broad topics are useful too



Areas of Application

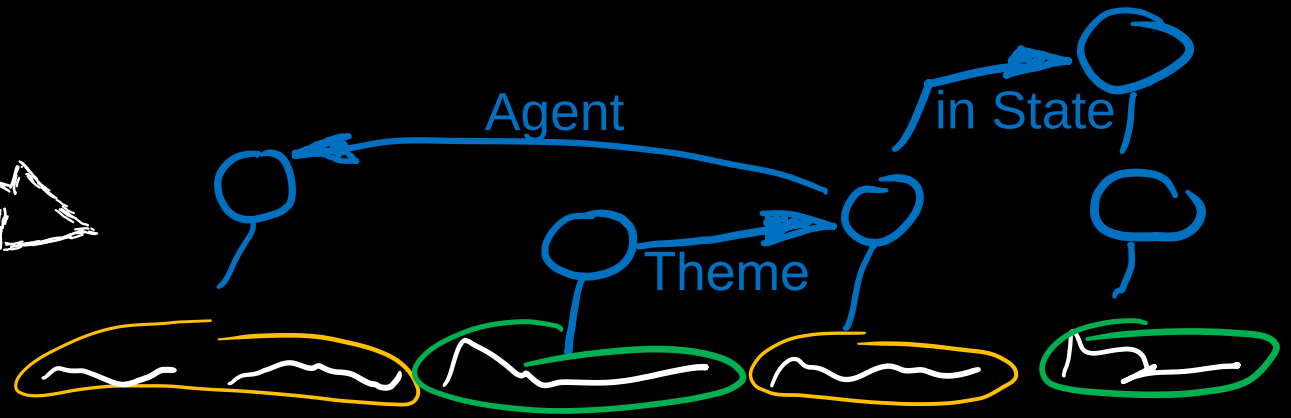
What's it for?

Areas of Application



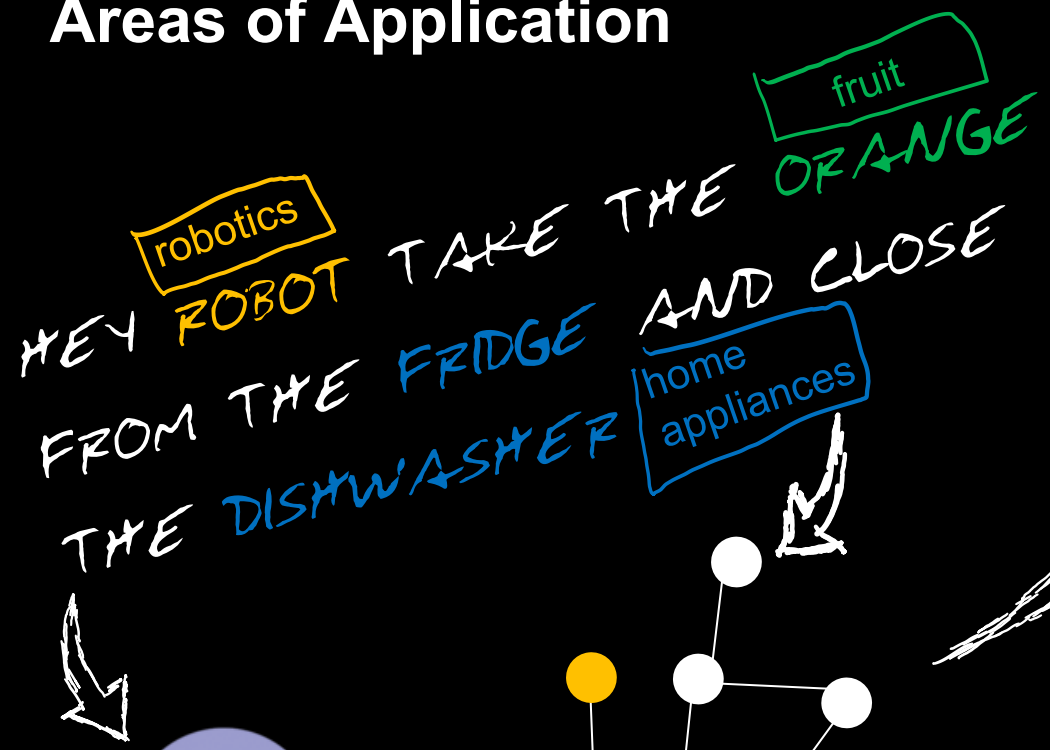
Cyc
(world knowledge)

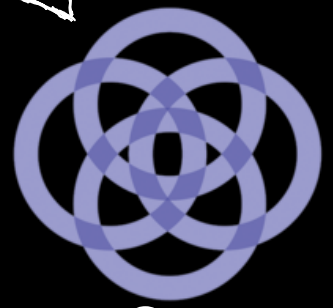
domain ontology



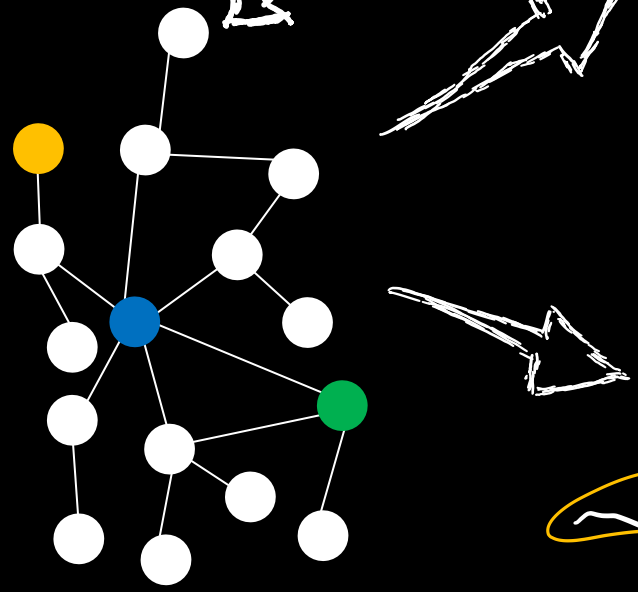
context modeling

Areas of Application



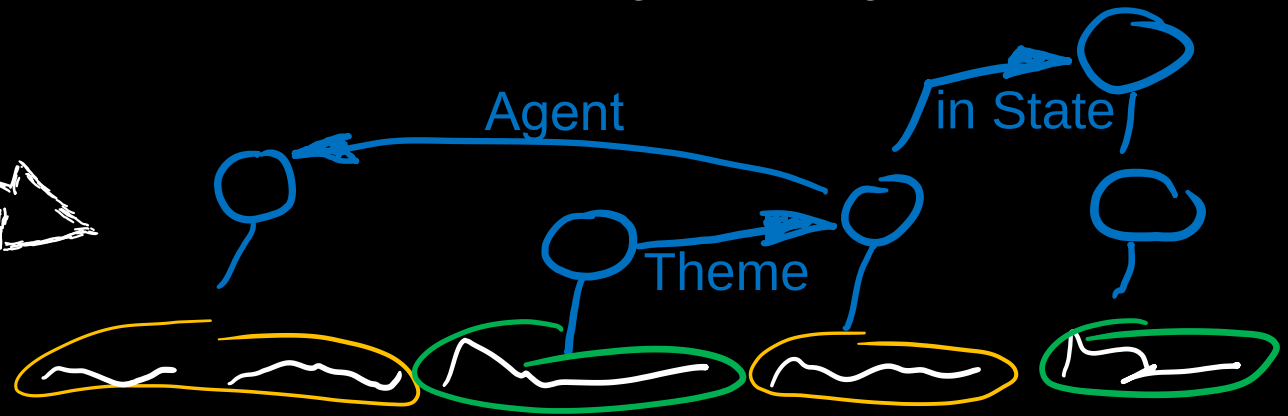


Cyc
(world knowledge)



domain ontology

dialog modeling



context modeling

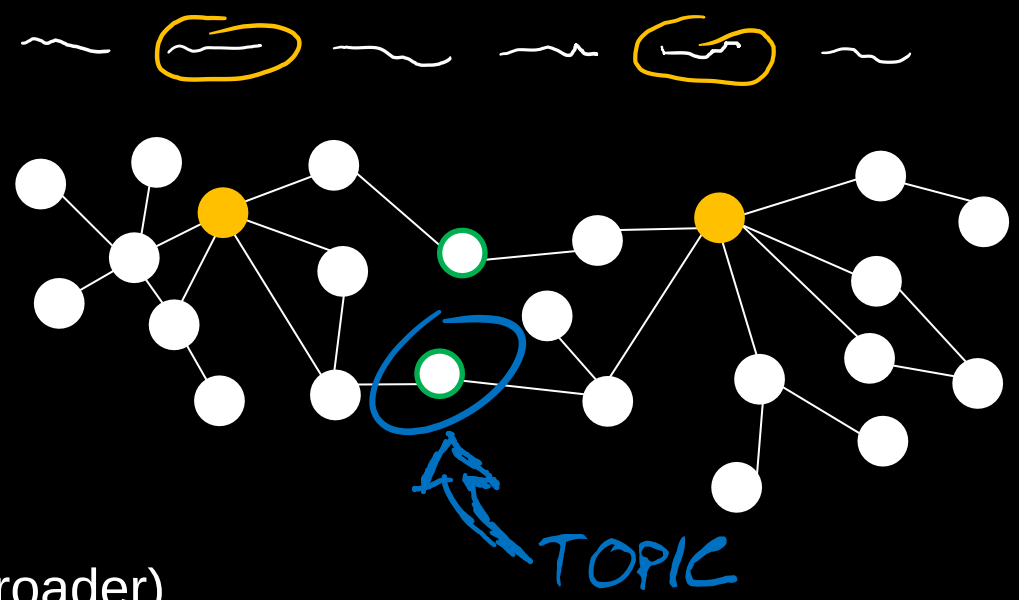
Summary

What to take away?

Summary

Unsupervised Multi-Topic Labeling for Spoken Utterances

- challenges: short, ungrammatical, erroneous
- approach:
  - disambiguate terms
  - build semantic graphs
  - determine most central concepts
    (two strategies: max vs. top)



- evaluation: 77.2% of the topics are a good fit (or broader)
  (user study with 6 subjects)
- applications: ontology selection, context and dialog modeling, …

References (1)

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, Mar. 2003.

[2] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, “Automatic Labeling of Topics,” in 2009 Ninth International Conference on Intelligent Systems Design and Applications, ser. ISDA ’09. Washington, DC, USA: IEEE Computer Society, Nov. 2009, pp. 1227–1232.

[3] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, “Unsupervised graph-based topic labelling using dbpedia,” in Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ser. WSDM ’13. New York, NY, USA: ACM, 2013, pp. 465–474. [Online]. Available: http://doi.acm.org/10.1145/2433396.2433454

[4] K. Coursey, R. Mihalcea, and W. Moen, “Using Encyclopedic Knowledge for Automatic Topic Identification,” in Proceedings of the Thirteenth Conference on Computational Natural Language Learning, ser. CoNLL ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 210–218.

[5] C. Cerisara, “Automatic discovery of topics and acoustic morphemes from speech,” Computer Speech & Language, vol. 23, no. 2, pp. 220–239, 2009.

[6] T. J. Hazen, M. Siu, H. Gish, S. Lowe, and A. Chan, “Topic modeling for spoken documents using only phonetic information,” in 2011 IEEE Workshop on Automatic Speech Recognition Understanding, Dec 2011, pp. 395–400.

References (2)

[7] M.-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, “Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery,” Computer Speech & Language, vol. 28, no. 1, pp. 210–223, 2014.

[8] R. Mihalcea, “Using Wikipedia for Automatic Word Sense Disambiguation,” in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 196–203.

[9] R. Mihalcea and A. Csomai, “Wikify!: Linking Documents to Encyclopedic Knowledge,” in Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ser. CIKM’07. New York, NY, USA: ACM, 2007, pp. 233–242.

[10] K. Coursey and R. Mihalcea, “Topic Identification Using Wikipedia Graph Centrality,” in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, ser. NAACL-Short ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 117–120.