The University Of Sheffield.

Semantics 2018 - 12 September 2018

Adapted TextRank for Term Extraction: A Generic Method of Improving Automatic Term Extraction Algorithms

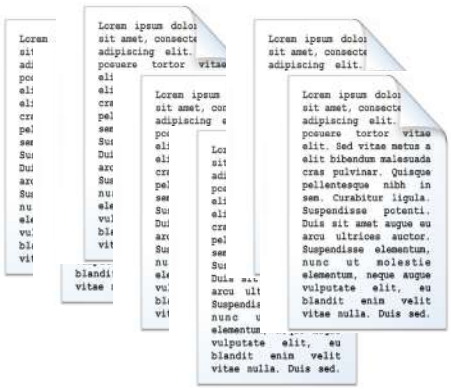
Ziqi Zhang[1], **Johann Petrak**[2], Diana Maynard[2]
ziqi.zhang@sheffield.ac.uk, johann.petrak@sheffield.ac.uk, d.maynard@sheffield.ac.uk

1. Information School, The University of Sheffield, UK
2. Department of Computer Science, The University of Sheffield, UK

The Task of ATE

- **Input:** (reasonably large) domain specific, focused corpus
- **Output:** list terms from the corpus, representing the domain
- **Approach**
  - Candidate extraction: domain-dependent, usually noun phrases, n-grams, or sequence matched by PoS patterns
  - **Candidate ranking & selection**: scoring candidates based on corpus statistics, selection by threshold, or machine learning

Domain specific corpus



ATE

Candidate Extraction

Candidate Ranking, Selection

Terms for the corpus

[semantic, 0.67,
ontology, 0.34,
nlp, 0.33
text mining, 0.12
…
web page, 0.012]

The University Of Sheffield.

The Task of ATE

- **A classic text mining problem**
  - Dating back to 1990s (Bourigault 1992)
  - To date still an active area of research
- **A fundamental step to many complex tasks**
  - Ontology engineering
  - Dictionary, terminology construction
  - Information Retrieval
  - Translation
  - …
- **Context of this work: KNOWMAK** (https://www.knowmak.eu/)

The Task of ATE

Differentiation from related tasks

		ATE
Keyword Extraction	- document specific - only a handful - mainly for indexing	- domain specific - # depends on corpus - mainly knowledge acquisition
NER	- usually real world named entities - sentence context is more important - semantic typing	- domain specific terms - corpus level statistics are more important - no typing

descriptions of electronic products from several
shops that offer Microdata markup. We prese
for each step of the data integration process i
crodata extraction, product classification, pro
extraction, identity resolution, and data fusion
ate our processing pipeline using 1.9 million p
from 9240 e-shops which we extracted from t
Crawl 2012, a large public Web corpus.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieva**
Analysis and Indexing; H.3.3 [**Information S**
Retrieval]: Information Search and Retrieva
tion Filtering

Keywords

Microdata; Information Extraction; Data Integ
tity Resolution

1. INTRODUCTION

Source: https://imanage.com/blog/named-entity-recognition-ravn-part-1/

The University Of Sheffield.

Motivation and Contribution

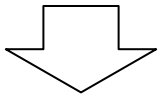
- **ATE still an unsolved problem**
  - No ‘all-rounder’ method
  - Performance always depends on data and domain
  - ‘one-size-fits-all’ solution feasible?
- **ATE methods are predominantly unsupervised**
  - For many domains there are already domain specific resources potentially useful, e.g., unlabelled corpus, pre-compiled named entity lists, partial ontologies, etc
  - Can we benefit from those?

The University Of Sheffield.

Motivation and Contribution

- **ATE still an unsolved problem**
  - No ‘all-rounder’ method
  - Performance always depends on data and domain
  - ‘one-size-fits-all’ solution feasible?

- **ATE methods are predominantly unsupervised**
  - For many domains there are already domain specific resources potentially useful, e.g., unlabelled corpus, pre-compiled named entity lists, partial ontologies, etc
  - Can we benefit from those?



A generic method that employs semantic relatedness to a set of **domain specific seed words** to potentially **improve any ATE** algorithms (by up to 25 percentage points in average precision in experiments).

AdaText - Overview

Adapted TextRank for Automatic Term Extraction

Domain specific seed words/ phrases

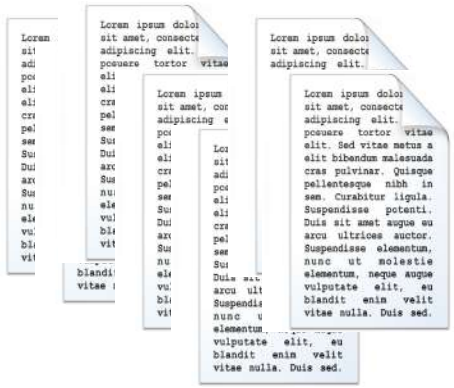
Semantic relatedness

Filter by threshold

TextRank

$[\ w_1=0.67,\ w_2=0.34,\ w_3=0.22,\ \ldots\ ]$

Extract words



Domain specific corpus

ATE (any algorithm)

$[\ t_1=1.99,\ t_2=1.21,\ t_3=1.10,\ \ldots\ ]$

+

Re-rank

$[\ t_1=2.19,\ t_3=1.41,\ t_2=1.29,\ \ldots\ ]$

The University Of Sheffield.

AdaText - Overview

Adapted TextRank for Automatic Term Extraction

Domain specific seed words/ phrases

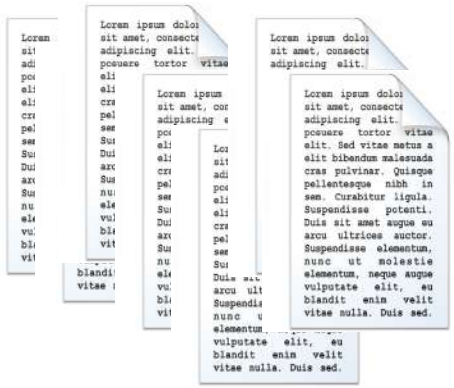
oolong, earl grey, cream, espresso, herbal, chamomile, pu-erh, americano, ginger, cappuccino, green, macchiato, vanilla, mocaccino, dessert, rooibus

Semantic relatedness

Filter by threshold

SEEDING

Extract words



Domain specific corpus

ATE (any algorithm)

CORPUS LEVEL TEXTRANK

TextRank

[w_1=0.67, w_2=0.34, w_3=0.22, …]

[t_1=1.99, t_2=1.21, t_3=1.10, …]

+

Re-rank

[t_1=2.19, t_3=1.41, t_2=1.29, …]

COMBINING WITH ATE

The University Of Sheffield.

AdaText - Seeding

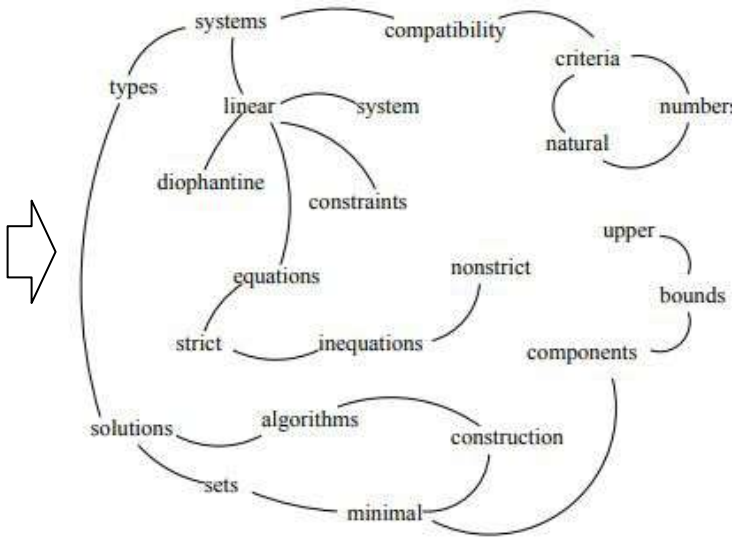
- **Input**
  - *C* - the target corpus from which terms are extracted
  - *S* - a set of 'seed' word/phrases representing the domain
    - taken from existing domain lexicons, or generated in an unsupervised way from available corpora
    - May not contain real terms from *C*
- **Process**
  - Extract words from *C*, as *W*
  - Compute pairwise semantic relatedness for *S x W*
    - Cosine similarity using GloVe embedding vectors
    - OOV ignored, phrase based on compositional averaging (Iyyer et al. 2015)
- **Output**
  - W_{sub} a subset of *W*, satisfying relatedness > ***min***
    Intuitively, they are more 'relevant' to the domain

The University Of Sheffield.

AdaText - Corpus Level TextRank

- **Input**
  - *C* - the target corpus from which terms are extracted
  - W_{sub} - the subset of words selected before
- **Process**
  - Apply TextRank to the graph created for W_{sub} to compute a TextRank (*tr*) score of every word *w* in W_{sub}
  - Traditional TextRank (Mihalcea et al., 2004) is a PageRank process to a graph of words from each document, where an edge is created if words co-occur in a context window of ***win***

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types



The University Of Sheffield.

AdaText - Corpus Level TextRank

- **Input**
  - *C* - the target corpus from which terms are extracted
  - W_{sub} - the subset of words selected before
- **Process**
  - Apply TextRank to the graph created for W_{sub} to compute a TextRank (*tr*) score of every word *w* in W_{sub}
  - Here it is adapted in two ways
    - A graph of words from the entire corpus
    - An edge is created if two words appear within ***win*** *anywhere in the corpus* (in any document)
- **Output**
  - *tr* scores for every word *w* in W_{sub}

AdaText - Combining with ATE

- **Input**
  - *C* - the target corpus from which terms are extracted
  - *ATE* - some ATE algorithm
  - *tr* scores for every word *w* in W_{sub}
- **Process**
  - Apply *ATE* to *C* to extract and score candidate terms
  - Revise each candidate term's score using *tr* scores for its composing words

$$score(t_i) = (1.0 + \frac{\sum_{w_i \in words(t_i)} tr(w_i)}{|words(t_i)|}) \times ate(t_i)$$

  - Then re-rank candidate terms by the new score
- **Output**
  - Re-ranked list of candidate terms

The University Of Sheffield.

Experiment and Findings

- **Base ATE methods** (as AdaText needs ATE scores of candidate terms)
  - Modified TFIDF (Zhang et al., 2016)
  - CValue (Ananiadou 1994)
  - Basic (Bordea et al., 2013)
  - RAKE (Rose et al., 2010)
  - Weirdness (Ahmad et al., 1999)
  - LinkProbability (LP, Astrakhantsev, 2016)
  - X^2 (Matsuo et al., 2003)
  - GlossEx (Park et al., 2002)
  - Positive Unlabelled (PU) learning (Astrakhantsev, 2016)
  - AvgRel - average relatedness score with seeds

- Use implementations:
  - JATE (https://github.com/ziqizhang/jate)
  - ATR4S (https://github.com/ispras/atr4s)

The University Of Sheffield.

Experiment and Findings

Evaluation measures

- Precision for top *K* ranked candidate terms
- *K* = {50, 100, 500, 1000, 2000}
- **Average P@K for all five *K*'s**

The University Of Sheffield.

Experiment and Findings

Datasets

- GENIA
  - 2,000 semantically annotated Medline abstracts
  - 434k words
  - 33k target terms
- ACLv2
  - 300 ACL paper abstracts
  - 32k words
  - 3k target terms

The University Of Sheffield.

Experiment and Findings

Seeds and parameters

- For GENIA:
  5,502 named entities from the BioNLP Shared Task 2011, only **25** match candidate terms

- For ACLv2:
  1,301 noun phrases from the titles of ACL, NAACL, and EACL papers (since 2000), ***none*** matches candidate terms

- Semantic relatedness threshold *min*=0.5 to 0.85 with 0.05 increment (selects for GENIA/ACL ~ 50/70 % … 10/5 %)

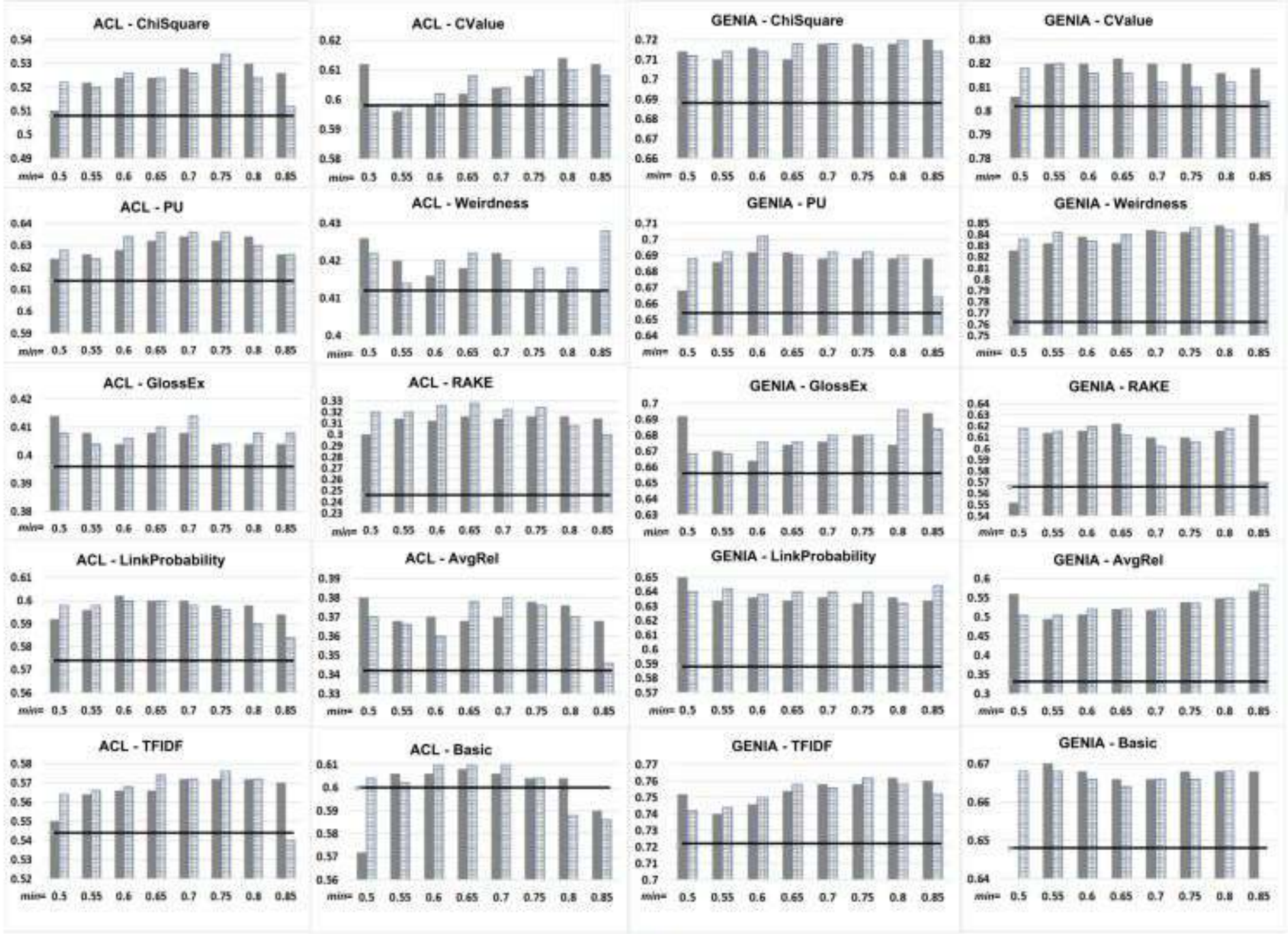
- TextRank context window *win*=5, 10

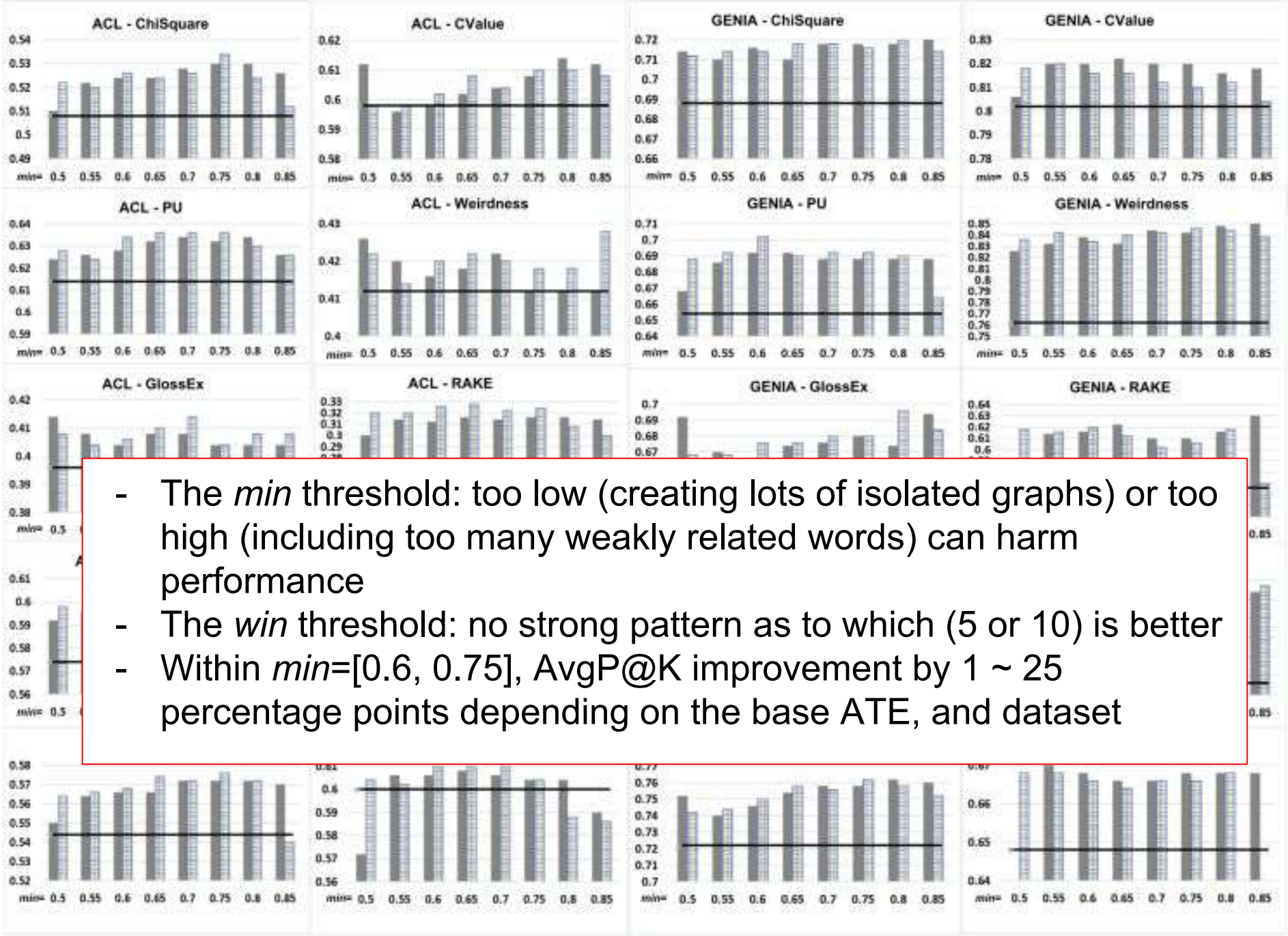
The University Of Sheffield.

Result - Base ATE

	Basic	LP	PU	CValue	GlossEx	RAKE	TFIDF	Weirdness	χ^2	AvgRel
					ACLv2					
P@50	**.84**	.72	.82	.62	.44	.18	.64	.40	.58	.30
P@100	.72	.69	**.82**	.69	.46	.15	.65	.50	.62	.34
P@500	.56	.56	.60	**.67**	.34	.29	.53	.36	.48	.39
P@1,000	.49	.51	.43	**.56**	.36	.29	.47	.40	.45	.35
P@2,000	.39	.39	.40	**.45**	.38	.32	.43	.40	.41	.33
AvgP@K	.60	.57	**.61**	.60	.40	.25	.54	.41	.52	.34
					GENIA					
P@50	.80	.38	.74	.86	**.88**	.68	.68	.78	.66	.16
P@100	.74	.51	.69	**.83**	.82	.63	.65	.74	.69	.28
P@500	.64	.70	.65	**.80**	.58	.56	.74	.78	.71	.36
P@1,000	.57	.69	.61	**.78**	.53	.52	.77	.77	.71	.44
P@2,000	.49	.66	.58	.74	.47	.44	**.77**	.74	.67	.42
AvgP@K	.65	.59	.65	**.80**	.66	.57	.72	.76	.69	.33

- Base ATE performance varies significantly depending on datasets.
- No single, consistently winning method on all five *K's*.
- E.g., PU is the best performing in **AvgP@K** on the ACL corpus, but the fourth worst performing on the GENIA corpus.


ACL - ChiSquare
ACL - CValue
GENIA - ChiSquare
GENIA - CValue
ACL - PU
ACL - Weirdness
GENIA - PU
GENIA - Weirdness
ACL - GlossEx
ACL - RAKE
GENIA - GlossEx
GENIA - RAKE
ACL - LinkProbability
ACL - AvgRel
GENIA - LinkProbability
GENIA - AvgRel
ACL - TFIDF
ACL - Basic
GENIA - TFIDF
GENIA - Basic
min= 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85
win=5
win=10



- The *min* threshold: too low (creating lots of isolated graphs) or too high (including too many weakly related words) can harm performance
- The *win* threshold: no strong pattern as to which (5 or 10) is better
- Within *min*=[0.6, 0.75], AvgP@K improvement by 1 ~ 25 percentage points depending on the base ATE, and dataset

The University Of Sheffield.

Conclusion

- **The takeaway message**
  - There is probably never a ‘one-size-fit-all’ ATE method, instead, think about improving existing ones
  - AdaText makes use of existing domain resources and builds on the TextRank algorithm
  - Generic method able to improve, potentially, any ATE method

- **Future work**
  - Whether and how the size and source of the seed lexicon affects performance
  - Adapt TextRank to a graph of both words and phrases, and see how this affects results

The University Of Sheffield.

Resources and Software

- **Data**
  - Genia corpus, ACL corpus available
  - Glove embeddings available
- **Software**
  - JATE (https://github.com/ziqizhang/jate)
  - ATR4S (https://github.com/ispras/atr4s)
  - Code for this work: https://github.com/ziqizhang/texpr
- **Slides**
  - https://goo.gl/1sPuhg

The University Of Sheffield.

References

1. Bourigault, D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In 14th International Conference on Computational Linguistics - COLING 92, 977–98
2. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daume, H., 2015. Deep unordered composition rivals syntactic methods for text classification, in: Association for Computational Linguistics. URL: docs/2015_acl_dan.pdf.
3. Mihalcea, R., Tarau, P., 2004. TextRank: Bringing order into texts, in: Proc. of EMNLP'04.
4. Zhang, Z., Gao, J., Ciravegna, F., 2016. Jate 2.0: Java automatic term extraction with apache solr, in: Proc. of LREC'16
5. Ananiadou, S., 1994. A methodology for automatic term recognition, in: Proc. of COLING1994, ACL, Stroudsburg, PA, USA. pp. 1034–1038.
6. Bordea, G., Buitelaar, P., Polajnar, T., 2013. Domain-independent term extraction through domain modelling, in: Proc. of the Conference on Terminology and Artificial Intelligence.
7. Astrakhantsev, N., 2015. Methods and software for terminology extraction from domainspecific text collection, in: Ph.D. thesis. Institute for System Programming of Russian Academy of Sciences.
8. Rose, S., Engel, D., Cramer, N., Cowley, W., 2010. Automatic keyword extraction from individual documents. John Wiley and Sons.
9. Ahmad, K., Gillam, L., Tostevin, L., 1999. University of surrey participation in trec 8: Weirdness indexing for logical document extrapolation and retrieval (wilder), in: Proc. of TREC1999.
10. Astrakhantsev, N., 2016. Atr4s: Toolkit with state-of-the-art automatic terms recognition methods in scala. arXiv preprint arXiv:1611.07804.
11. Matsuo, Y., Ishizuka, M., 2003. Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13, 157–169.
12. Park, Y., Byrd, R., Boguraev, B., 2002. Automatic glossary extraction: Beyond terminology identification, in: Proc. of COLING'02, Association for Computational Linguistics. pp. 1–7.

The University Of Sheffield.

Acknowledgements

This work is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992 (KNOWMAK project)

https://www.knowmak.eu/

The University Of Sheffield.

Thank you!