Complex word networks - comparing and combining information extraction methods

Keyword extraction – purposes

- Overview over content
- Classification of content
- Monitoring of thematic trends
- Initial step of network based topic modelling

Keyword extraction – methods

- Linguistic methods
  - ☹ Depends on language
- Supervised machine learning methods
  - ☹ Requires labelled training sets
  - ☹ Problematic for discovering new trends
- Unsupervised methods
  - ☺ Can be used on huge corpora without "manual" work
  - Word frequencies
    - Simplistic: Count words/phrases: **Tf**
    - Better: Higher weight if term is unusual for corpus: **Tf-Idf** (Spärck Jones, K., 1972)
  - Information / entropies
    - **Entropy**: Partition text into P parts and use $S(w) \propto -\sum_1^P Tf_i(w)\ \ln(\sum_1^P Tf_i(w))$ for measuring the heterogeneity of the distribution of w. (P. Herrera, J., Pury, P., 2008)
    - **Tsallis Entropy**: Use non-additive $S_q(w) \propto \frac{1-(Tf_i(w))^q}{q+1}$ for measuring heterogeneity and finding optimal q.(Mehri, A., Darooneh A. H., 2011)
  - Graph-based arguments
    - Apply PageRank algorithm on network where nodes are words and edges indicate neighbourhood
    - First version called **TextRank** (Mihalcea, R., Tarau, P., 2004)
    - Many variations like **TopicRank** (Bougouin, A., Boudinand, F., Daille, B., 2013) and **PositionRank** (Florescu, C., Caragea, C., 2017)

All unsupervised methods rest on three observations

- **Keywords frequent in document but not too frequent in corpus**
- **Keywords not homogeneously distributed in a document but clustered**
- **Keywords more likely near the beginning of a document**

➔Find a method which exploits all three observations

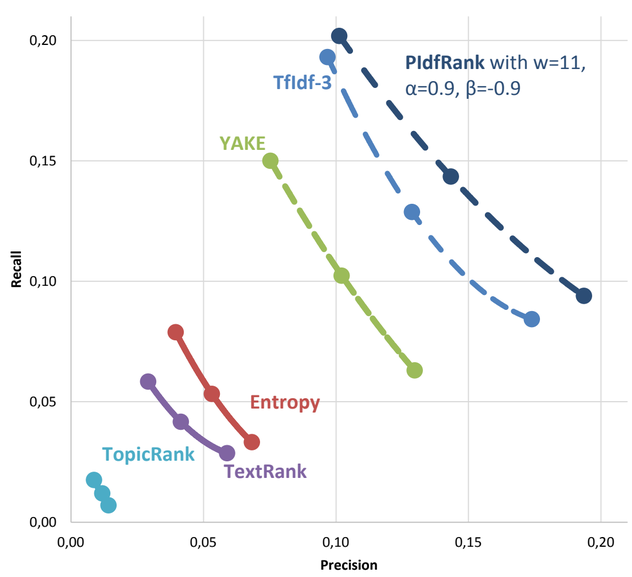
New variant of graph-based rank method:

"Positional IdfRank"

Includes Idf and position into the word network.
See example at right hand side.

Compare various methods regarding recall and precision, using the keyword-labelled corpus SemEval 2010 (244 CS papers).
But this way of measuring the keyword extraction quality is problematic: Human assigned keywords are not necessarily the best, nor even good.

Word (not phrase) recall vs precision for various extraction methods

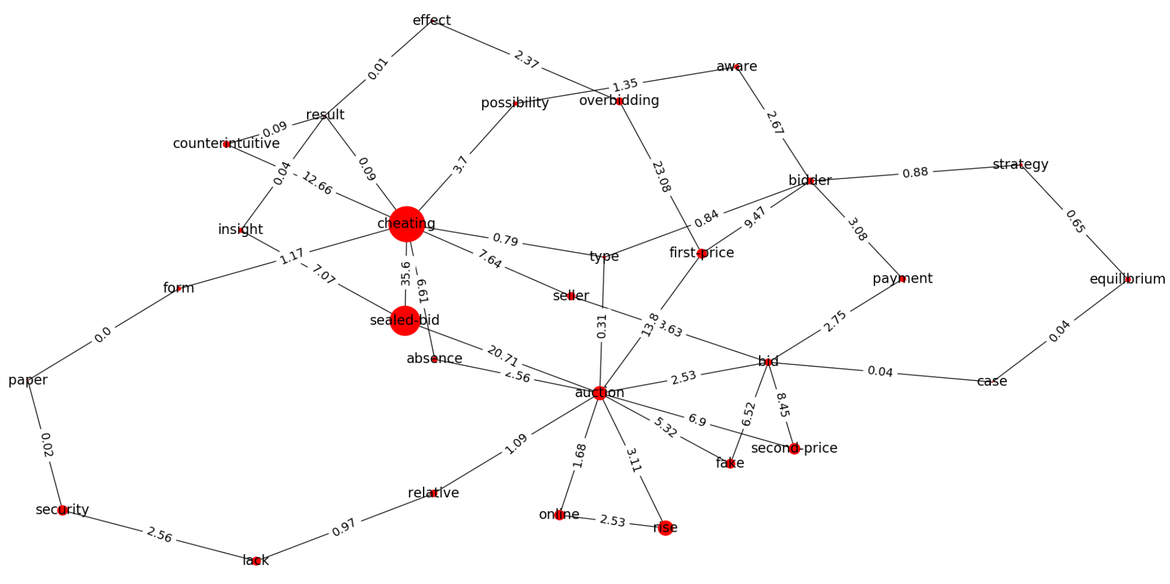


A toy example – title and abstract of one of the documents of the SemEval 2010 corpus

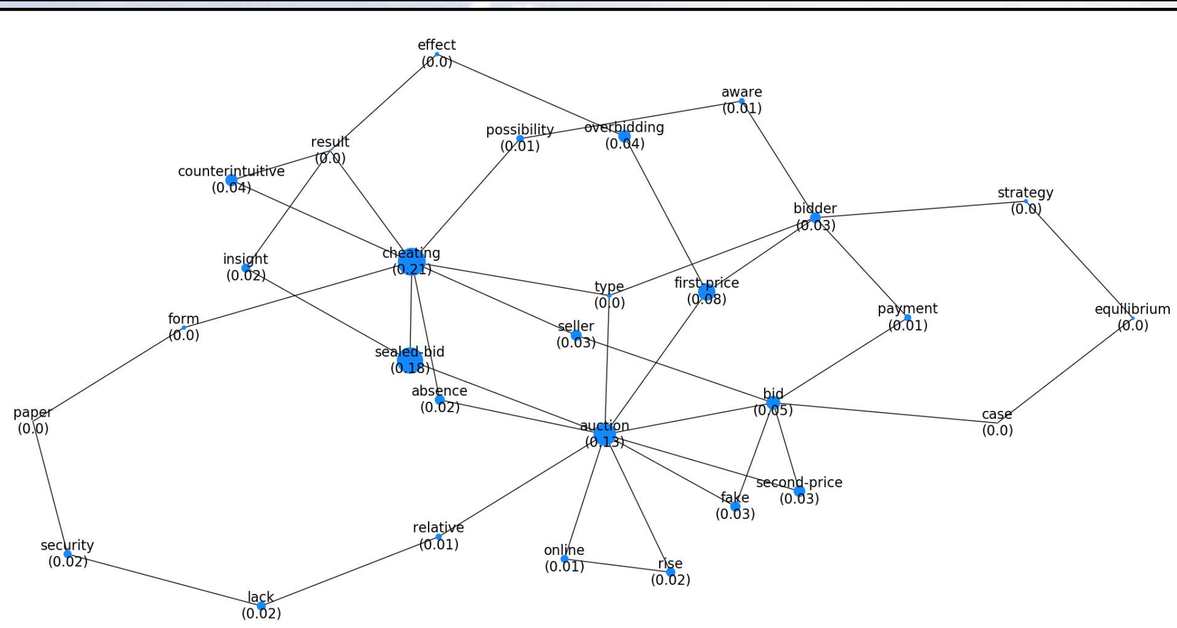
On Cheating in Sealed-Bid Auctions

Motivated by the rise of online auctions and their relative lack of security, this paper analyzes two forms of cheating in sealed-bid auctions. The first type of cheating we consider occurs when the seller spies on the bids of a second-price auction and then inserts a fake bid in order to increase the payment of the winning bidder. In the second type, a bidder cheats in a first-price auction by examining the competing bids before deciding on his own bid. In both cases, we derive equilibrium strategies when bidders are aware of the possibility of cheating. These results provide insights into sealed-bid auctions even in the absence of cheating, including some counterintuitive results on the effects of overbidding in a first-price auction.
From: Porter, R., Shoham, Y. (2004). Decision Support Systems. 39.

1st step: Preparation – Remove punctuation; keep only nouns and adjectives; lemmatise
2nd step: Graph building - Every remaining word is a node
Two nodes are linked if they are in common window of width w
Edge weight $w_{ij} = Pf_{ij} * Idf_i * Idf_j$, where Pf_{ij} is pair frequency
3rd step: Preferences – Preference for choosing node randomly: $p_i = (1 + pos_i)^{\beta} * Idf_i$, $\beta \leq 0$.



4th step: Simulate "random reader" – Markov chain $x_{t+1} = Gx_t$ with
$G_{ij} = \alpha\widetilde{w}_{ij} + (1-\alpha)p_i\delta_{ij}$ where $\widetilde{w}_{ij}$is row-normalised w_{ij}
Stationary distribution determines node ranking (Page, L., Brin, S., 1998)



Top keywords: 'cheating', 'sealed-bid', 'auction', 'first-price', 'bid', 'overbidding', 'counterintuitive', 'second-price', 'seller', 'bidder'

From word scores $s_i$calculate scores for n-grams: $s_{(ijk)} = \sum_{\gamma\in\{i,j,k\}} \frac{tf_{ijk}}{tf_\gamma} s_\gamma$

Top keyphrases: 'cheating sealed-bid auction', 'first-price auction', 'form cheating sealed-bid'

Another example: *Realistic Cognitive Load Modeling for Enhancing SharedMental Models in Human-Agent Collaboration*, Xiaocong Fan

TfIdf3	PIdfRank3	TextRank	Entropy	TopicRank	PositionRank	YAKE
cognitive	cognitive	loaded	team	mental	human cognitive load	agent
team	cognitive load	model	state	evolution	cognitive load model	team
load	team	state	information	human	realistic cognitive load	load
cognitive load	mental	information effectiveness	task	state current time step function	current load agent	human
agent	human	timely	agent	human team member expectation	human partner agent	model
mental	realistic	human team member expectation	human	load effect	human agent pair	cognitive
human	mental model	performative	mmop	enhancing	load agent	information
mental model	modeling enhancing shared	communication	load	human bias agent model human	human team member	cognitive load
secondary task	enhancing shared human	agent pair hap	cognitive	cognitive resource task human	human agent collaboration	task
teammate	load modeling enhancing	capacity	th	mouse	cognitive agent architecture	mental model
processing load	teammate	realistic cognitive	info	exact order	instantaneous cognitive load	mental
team member	agent	mental	secondary	realistic cognitive load	short human agent	processing
multi party	team member	info	party	info teammate process half info	human agent	performance
info	secondary task	teammate	teammate	receiver belief party	cognitive load study	multi party
load model	human partner	step	type	multi party	cognitive load theory	secondary task
model	info	processing	communication	behavior paper concept	th2<th3<th1 agent team	time
human partner	processing load	multi	step	memory	relative cognitive load	processing load
capacity	capacity	individual group task	mental	popup	cognitive load it"s	team member
information	cognitive capacity	measurement	performance	information sharing communications haps	resistant cognitive load	party

Stimulating discussions and productive collaboration with Mark Azzam, Rasmus Beckmann, Simon Odrowski, and Jana Thelen are gratefully acknowledged.

Andreas Hamm
Think Tank
German Aerospace Center (DLR), Cologne, Germany
Andreas.Hamm@dlr.de

Deutsches Zentrum für Luft- und Raumfahrt