Hypergraphs and Information Fusion for Term Representation Enrichment. Applications to Named Entity Recognition and Word Sense Disambiguation

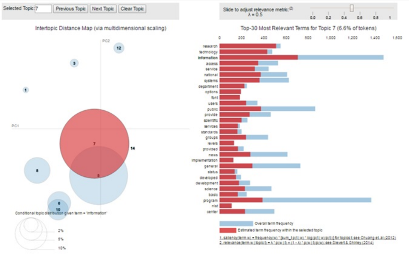
Ph.D. Thesis Defense

Pavel Soriano-Morales

Supervised by Sabine Loudcher and Julien Ah-Pine

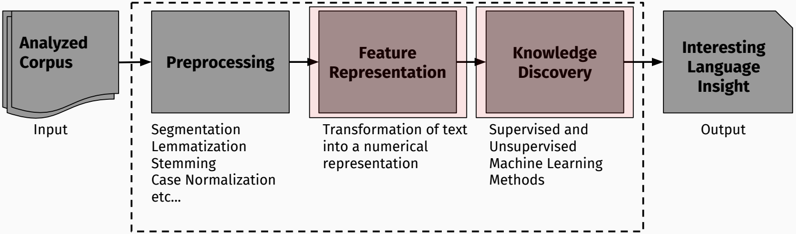
February 7th, 2018

Why it is useful to us to understand text?



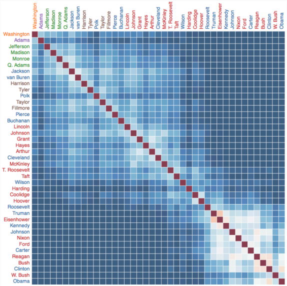
How do we extract meaning from text?

We use **Natural Language Processing** (NLP), a field of computer science interested on making computers extract useful information from text

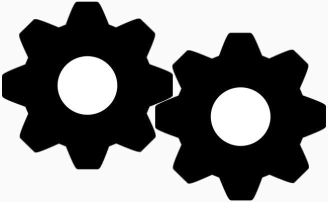


Feature Representation and Knowledge Discovery

How do we represent text for the machine to understand?



What techniques do we use to discover meaning from text?



Representing Text

- **Three common ways to represent text**

Representing Text

- **Three common ways to represent text**
  - Lexical

Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic

Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree

Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree

Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
- **Working Example**

Representing Text

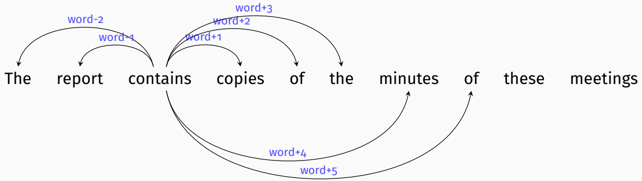
- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
- **Working Example**

  The report contains copies of the minutes of these meetings

Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
- **Working Example**

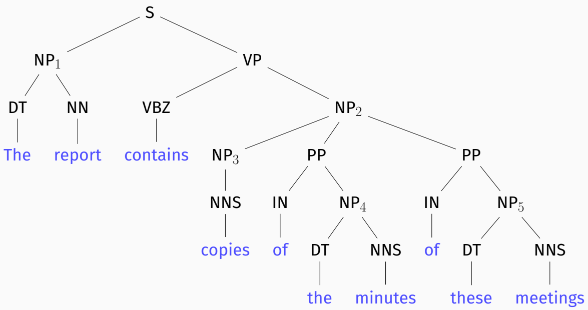
  The report contains copies of the minutes of these meetings



Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
- **Working Example**

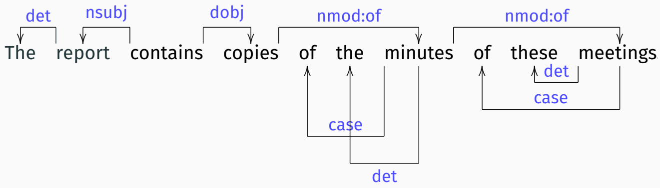
  The report contains copies of the minutes of these meetings



Representing Text

- **Three common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
- **Working Example**

  The report contains copies of the minutes of these meetings

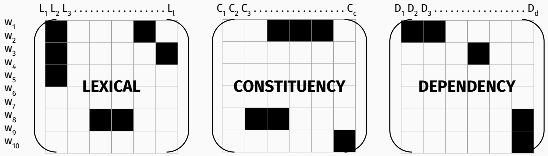


Representing Text

- **Text Representation Models**
  - Words and features can be represented by means of graph-based models matrices
  - Or directly with (sparse) matrices
- **Leveraging the Network Structure**
  - We can find communities of similiar words according to their meaning

Representing Text

- **Text Representation Models**
  - Words and features can be represented by means of graph-based models matrices
  - Or directly with (sparse) matrices
- **Leveraging the Network Structure**
  - We can find communities of similiar words according to their meaning



Main Challenges and Contributions

1. What type of model can we employ to represent a corpus using heterogeneous features?

Main Challenges and Contributions

1. What type of model can we employ to represent a corpus using heterogeneous features?
    - *Hypergraph linguistic model to hold different types of linguistic information*

Main Challenges and Contributions

1. What type of model can we employ to represent a corpus using heterogeneous features?
    - *Hypergraph linguistic model to hold different types of linguistic information*
2. How can we combine these features while dealing with feature sparsity?

Main Challenges and Contributions

1. What type of model can we employ to represent a corpus using heterogeneous features?
   - *Hypergraph linguistic model to hold different types of linguistic information*
2. How can we combine these features while dealing with feature sparsity?
   - *Multimedia fusion techniques to combine and densify representation spaces*

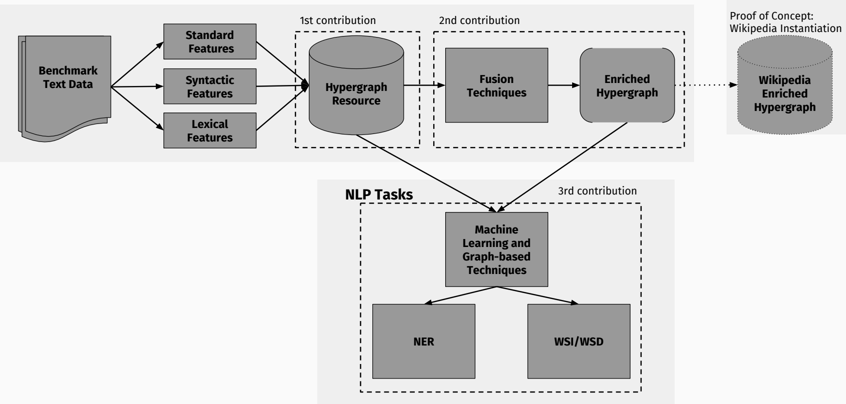
Main Challenges and Contributions

1. What type of model can we employ to represent a corpus using heterogeneous features?
   - *Hypergraph linguistic model to hold different types of linguistic information*
2. How can we combine these features while dealing with feature sparsity?
   - *Multimedia fusion techniques to combine and densify representation spaces*
3. How can we find and employ communities existing within the language networks?

Main Challenges and Contributions

1. What type of model can we employ to represent a corpus using heterogeneous features?
   - *Hypergraph linguistic model to hold different types of linguistic information*
2. How can we combine these features while dealing with feature sparsity?
   - *Multimedia fusion techniques to combine and densify representation spaces*
3. How can we find and employ communities existing within the language networks?
   - *An alternative network-based algorithm to discover semantically related words within a text*

Work Overview



Contributions in Detail

Hypergraph Linguistic Model

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**
  - Network Models [MTF04]

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**
  - Network Models [MTF04]
  - Vector Space Models [MS+99]

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**
    - Network Models [MTF04]
    - Vector Space Models [MS+99]

- **We choose network models**

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**
  - Network Models [MTF04]
  - Vector Space Models [MS+99]
- **We choose network models**
  - Used in a large quantity of NLP tasks [MR11]

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**
  - Network Models [MTF04]
  - Vector Space Models [MS+99]

- **We choose network models**
  - Used in a large quantity of NLP tasks [MR11]
  - Graphs structures can give us a clearer view into the relations of words within a text [CM09]

Introduction

Based on the distributional hypothesis, a word is defined by its surroundings, we can extract useful information from a text.

- **How do we represent textual data?**
  - Network Models [MTF04]
  - Vector Space Models [MS+99]

- **We choose network models**
  - Used in a large quantity of NLP tasks [MR11]
  - Graphs structures can give us a clearer view into the relations of words within a text [CM09]
  - Ultimately graphs are transformed to a vectorial representation through the adjacency/incidence matrices

Existing Language Networks

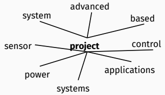
The report contains copies of the minutes of these meetings

Lexical Networks

Existing Language Networks

The report contains copies of the minutes of these meetings

Lexical Networks

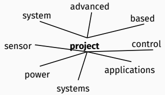


Syntactic Networks

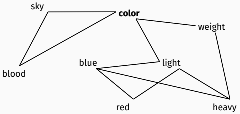
Existing Language Networks

The report contains copies of the minutes of these meetings

Lexical Networks



Syntactic Networks

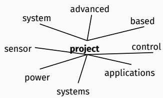


Semantic Networks

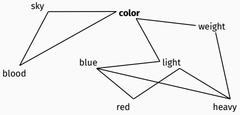
Existing Language Networks

The report contains copies of the minutes of these meetings

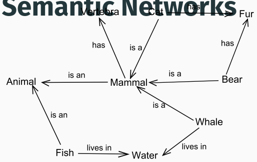
Lexical Networks



Syntactic Networks



Semantic Networks



An expert is usually involved.

Limitations and Proposition

- **Limitations of existing representations**

Limitations and Proposition

- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information

Limitations and Proposition

- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network may relate maximum two words at each time

Limitations and Proposition

- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network may relate maximum two words at each time
- **Proposition**

Limitations and Proposition

- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network may relate maximum two words at each time
- **Proposition**
  - Represent together linguistic co-occurrences through a hypergraph model

Limitations and Proposition

- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network may relate maximum two words at each time
- **Proposition**
  - Represent together linguistic co-occurrences through a hypergraph model
    - Link together three different types of networks, using lexical and syntactic data

Limitations and Proposition

- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network may relate maximum two words at each time
- **Proposition**
  - Represent together linguistic co-occurrences through a hypergraph model
    - Link together three different types of networks, using lexical and syntactic data
    - Get a semantic overview at three different levels: short range (with dependency functions), medium range (phrase constituency membership), and long range (lexical co-occurrence)

Proposed Model

- Explain (grpahically/with the working exampleh) we use lexical and syntactic info and the build a fusion of them with a hypergraph.

Contributions in Detail

Combining Features and Dealing with Sparsity

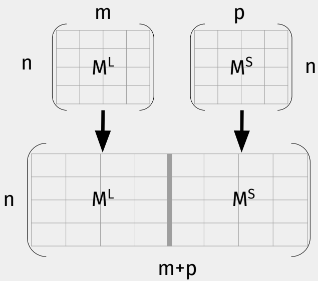
Multimedia Fusion Techniques [Atr+10; ABL10]:

- **Definition**
  - Set of techniques used in multimedia analysis tasks to integrate multiple media
  - The goal is to obtain rich insights about the data being treated
  - We adapt these techniques to our use case: textual information
- **Main fusion operators:**
  - Early Fusion $E_\alpha(\cdot)$,
  - Late Fusion $L_\beta(\cdot)$,
  - Cross Fusion $X_\gamma(\cdot), X_F(\cdot)$
  - α and β: Assign an importance weight to each of their operators
  - γ: number of top similar items to take from the similarity space

Early and Late Fusion

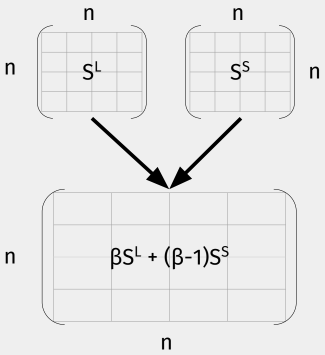
EARLY FUSION

Matrices M^L and M^S have the same number of rows

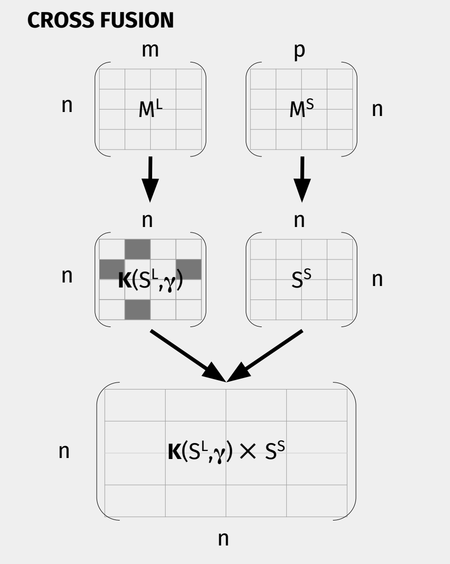


LATE FUSION: SIMILARITY FUSION

Matrices S^L and S^B have the same size



Cross Fusion



Hybrid Fusion 1

Put here some very visual way of representing hybrid fusion.

In fact, early and late fusion should be presented with the working example also.

Hybrid Fusion 2

Put here some very visual way of representing hybrid fusion.

In fact, early and late fusion should be presented with the working example also.

Leveraging the network communities

1. Show a large (with more text than that of my example) image of the hypergraph model

Contributions in Detail

Finding Communities in the Network

Leveraging the network communities 1

1. Link some words together with a color overlay to represent possible communities (clusters/groups) of same sense words.
2. Argue that thanks to the heterogeneous info contained in the structure, we can relate words according to different linguistic properties

Leveraging the network communities 2

1. Link some words together with a color overlay to represent possible communities (clusters/groups) of same sense words.
2. Argue that thanks to the heterogeneous info contained in the structure, we can relate words according to different linguistic properties

Applications to NLP

Hypergraph Model Instantiation

Applications

- We instantiate our proposed linguistic resource
  - Based on the English Wikipedia corpus
- Use the proposed model to solve two NLP tasks:
  - Named Entity Recognition
  - Word Sense Induction and Disambiguation
- These experiments have two main objectives:
  - Test the effectiveness of fusion enriched representations (heterogeneity + less sparse spaces)
  - Leverage the structure of the network built following our proposed model

Hypergraph Model Instantiation

- Introduction to SAEWD
- Motivation
- Characteristics
- Show small diagram of the process

Hypergraph Model Instantiation

- Image with how the hypergraph corpus is stored in files and how we can access the information via key-value pairs to select nouns or verbs or types of noun phrases etc

Wikipedia Feature Enriched Spaces

	Lexical Features (5.49%) M^L	**Syntactic Features (4.97%)** M^S	**Early Fusion (5.23%)** $E(M^L, M^S)$	X_F **Fusion (16.75%)** $X_F(S^S, M^L)$	X_F **Fusion (13.45%)** $X_F(S^L, M^S)$
	priests	monk	sailor	vassal	sailor
	nun	regent	regent	regent	fluent
priest	canton	aedile	nuclei	nun	dean
	sailor	seer	nun	sailor	nuclei
	burial	meek	relic	monk	chorus

Applications to NLP

Solving Named Entity Recognition

Definition and Objectives

- The goal is to automatically discover mentions that belong to a well-defined semantic category.
- The classic task of NER involves detecting among four types of entities and a non-entity class:
  - Location (LOC)
  - Organization (ORG)
  - Person (PER)
  - Miscellaneous (MISC)
  - None (O)
- We assess the effectiveness of the classic fusion methods and propose new hybrid combinations
- ** Show here graphical presentation of entities**

Representation Spaces

Lexical Space (L)

Word	Features
Australian	word:Australian, word+1:scientist, word+2:discovers
scientist	word-1:Australian, word:scientist, word+1:discovers, word+2:star
discovers	word-2:Australian, word-1:scientist, . . . , word+2:telescope
star	word-2:scientist, word-1:discovers, word:star, . . . , word+2:telescope
with	word-2:discovers, word-1:star, word:with, word+1:telescope
telescope	word-2:star, word-1:with, word:telescope

Syntactic Space (S)

Word	Contexts
Australian	scientist/NN/amod_inv
scientist	Australian/JJ/amod, discovers/VBZ/nsubj_inv
discovers	scientist/NN/nsubj, star/NN/dobj, telescope/NN/nmod:with
star	discovers/VBZ/dobj_inv
telescope	discovers/VBZ/nmod:with_inv

Standard Features Space (T)

- Each word
- Whether it is capitalized
- Prefix and suffix (of each word their surroundings)
- Part of Speech tag

Experimental Protocol

- **Preprocessing**
  - Normalize numbers
- **Test Corpora**
  - CoNLL-2003 (CONLL) [SM03]: Train: 219,554 lines. Test: 50,350
  - Wikiner (WNER) [NMC09]: No Train/Test split. 3.5 million words. Evaluated in a 5-fold CV
  - Wikigold (WGLD) [Bal+09]: No Train/Test split. 41,011 words. Evaluated in a 5-fold CV
- **Annotation Scheme**
  - **B**eginning, **I**nside, **O**utside
- **Learning Algorithm**
  - Structured Perceptron [Col02]
- **Evaluation Metrics**
  - Precision, Recall, F-measure

Evaluation

- Best Fusion operators on the F-measure over the three datasets.
- Achieved using a higher Degree fusion operator
- Notice the comparison with the Early Fusion baseline
- Visually show the best fusion operator, not with the formula.

		Triple Early Double Late Cross Feature Fusion ($EEELX_FLX_F$)		
		CONLL	WNER	WGLD
M^L	$\hat{b}_{EEELX_FLX_F}$	65.01	78.02	62.34
$M^L_{\alpha=0.95}$	$\hat{b}_{EEELX_FLX_F}$	**79.67**	**81.79**	**67.05**
EF Baseline		78.90	80.04	63.20

Analyzing the Best Fusion Operator

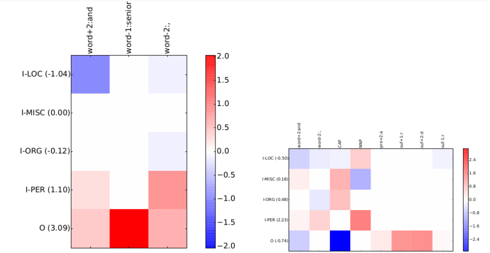
Decompose best fusion in four models:

$$\overbrace{\underbrace{\overbrace{E_{\alpha=0.95}(\underbrace{M^L}_{1}, M^T}^{2}, L(M^T, X_F(S^S, M^T))}_{3}, L(M^L, X_F(S^S, M^L))}^{4})$$

1. M^L used to train model M_1.
2. $E(\alpha_1 M^L, \alpha_2 M^T)$ used to train model M_2, with $\alpha_1 = 0.95, \alpha_2 = 0.05$
3. $E_\alpha(\alpha_1 M^L, \alpha_2 M^T, \alpha_3 L(M^T, X_F(S^S, M^T)))$ used to train model M_3, with $\alpha_1 = 0.95, \alpha_2 = \alpha_3 = 0.05$
4. $E_\alpha(\alpha_1 M^L, \alpha_2 M^T, \alpha_3 L(M^T, X_F(S^S, M^T)), \alpha_4 L(M^L, X_F(S^S, M^L)))$ used to train model M_4, with $\alpha_1 = 0.95, \alpha_2 = \alpha_3 = \alpha_4 = 0.05$

Analyzing the Best Fusion Operator

We focus on the word *Kory*, and its performance from model M_1 to M_2

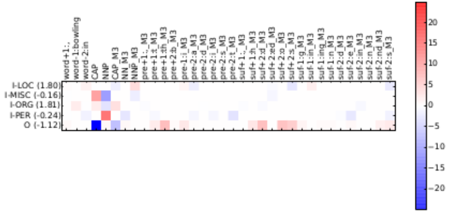


(a) The word *Kory* is predicted to be class O by model M_1 given the coefficients of the features indicated.

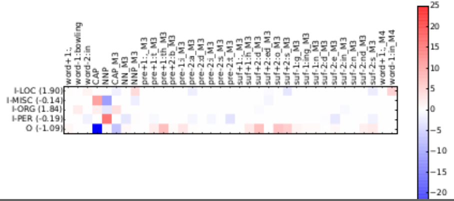
(b) The same word is correctly predicted to be class PER by model M_2 given the new features in the model.

Analyzing the Best Fusion Operator

We focus on the word *Green*, and its performance from model M_3 to M_4



(a) The word *Green* is predicted to be class ORG by model M_3 given the coefficients of the features indicated.



Applications to NLP

Solving Word Sense Induction and Disambiguation

Introduction

- Introduction to WSI/WSD

Experimental Protocol

- **Preprocessing**
  - Normalize numbers
- **Test Corpora**
  - Semeval 2007 [SM03]: Train: 219,554 lines. Test: 50,350
- **Clustering Algorithm**
  - Spectral Clustering
- **Evaluation Metrics**
  - Supervised: F-score
  - Unsupervised: Recall
  - Proposed: H-score

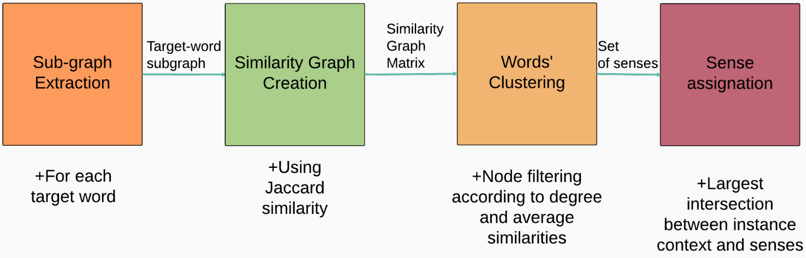
WSI/WSD: Evaluation

- Results for WSI/WSD with spectral clustering

How to exploit a linguistic network to solve word sense induction and disambiguation?

- **Similar approaches**
  - Hyperlex [Vé04]
  - University of York (UoY) [KM07]
- **Limitations of existing approaches**
  - Single typed networks
  - Large number of parameters
- **Features**
  - Be able to exploit different types of linguistic information (lexical or syntactic co-occurrence)
  - Keep the number of parameters low and allow for their automatic adjusting according to the network's nature

Proposed Method



Semeval Results

- Semeval 2007 results table

WSI/WSD: Evaluation

- Verbs and nouns behaviors
- Insight into senses found by the algorithm

Conclusions

Conclusions

Insights From our Contributions

Future Work

References I

References

References

Christopher D Manning, Hinrich Schütze, et al. *Foundations of statistical natural language processing.* Vol. 999. MIT Press, 1999.

References II

Michael Collins. "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms". In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1–8. DOI: 10.3115/1118693.1118694.

Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *CoNLL*. ACL, 2003, pp. 142–147.

References III

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. “PageRank on Semantic Networks, with Application to Word Sense Disambiguation”. In: *Proceedings of the 20th International Conference on Computational Linguistics*. COLING ’04. Geneva, Switzerland: Association for Computational Linguistics, 2004. DOI: 10.3115/1220355.1220517.

Jean Véronis. “HyperLex: lexical cartography for information retrieval”. In: *Computer Speech & Language* 18.3 (2004), pp. 223 –252. ISSN: 0885-2308. DOI: 10.1016/j.csl.2004.05.002.

References IV

Ioannis P. Klapaftis and Suresh Manandhar. “UOY: A Hypergraph Model for Word Sense Induction & Disambiguation”. In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. SemEval ’07. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 414–417.

Dominic Balasuriya et al. “Named Entity Recognition in Wikipedia”. In: *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*. People’s Web ’09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 10–18. ISBN: 978-1-932432-55-8. URL: http://dl.acm.org/citation.cfm?id=1699765.1699767.

References V

Monojit Choudhury and Animesh Mukherjee. “The Structure and Dynamics of Linguistic Networks”. English. In: *Dynamics On and Of Complex Networks*. Ed. by Niloy Ganguly, Andreas Deutsch, and Animesh Mukherjee. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser Boston, 2009, pp. 145–166. ISBN: 978-0-8176-4750-6. DOI: 10.1007/978-0-8176-4751-3_9.

Joel Nothman, Tara Murphy, and James R. Curran. “Analysing Wikipedia and Gold-standard Corpora for NER Training”. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL ’09. Athens, Greece: Association for Computational Linguistics, 2009, pp. 612–620.

References VI

Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. "Link communities reveal multiscale complexity in networks". In: *Nature* 466.7307 (2010), pp. 761–764.

Pradeep K. Atrey et al. "Multimodal fusion for multimedia analysis: a survey". In: *Multimedia Syst.* 16.6 (2010), pp. 345–379.

Rada F. Mihalcea and Dragomir R. Radev. *Graph-based Natural Language Processing and Information Retrieval.* 1st. New York, NY, USA: Cambridge University Press, 2011. ISBN: 0521896134, 9780521896139.