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

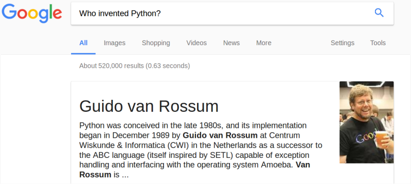
Ph.D. Thesis Defense

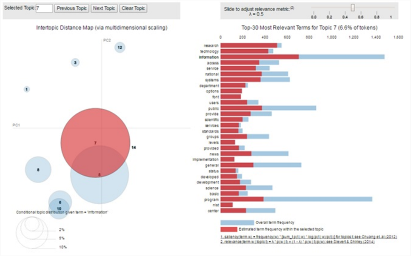
Pavel Soriano-Morales

Supervised by Sabine Loudcher and Julien Ah-Pine

February 7th, 2018

Why it is useful to us to automatically understand written language?

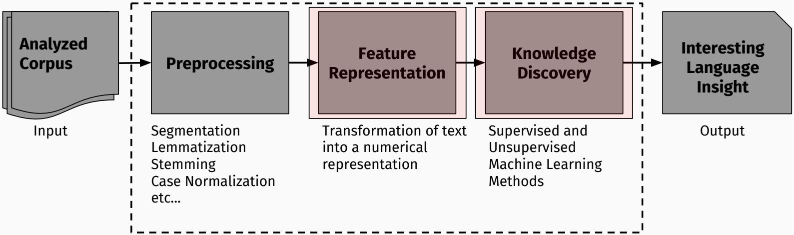




Introduction

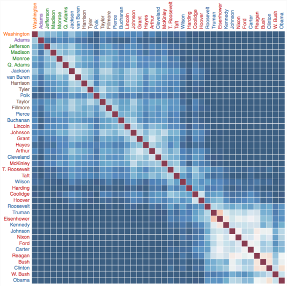
How do we extract meaning from written language?

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

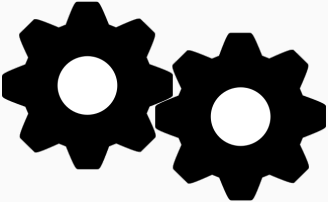


In this thesis, we focus on Feature Representation and Knowledge Discovery

How do we represent text for the machine to understand?



What techniques do we use to discover meaning from text?



Introduction

Example Phrase

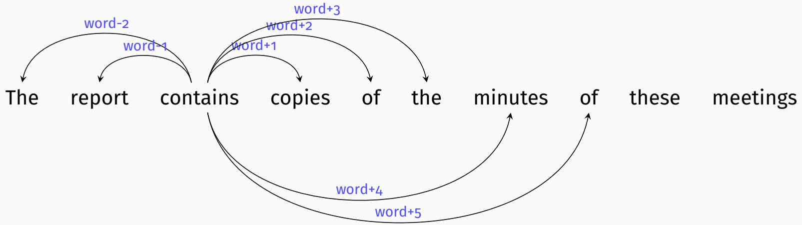
The report contains copies of the minutes of these meetings

Introduction

Example Phrase

The report contains copies of the minutes of these meetings

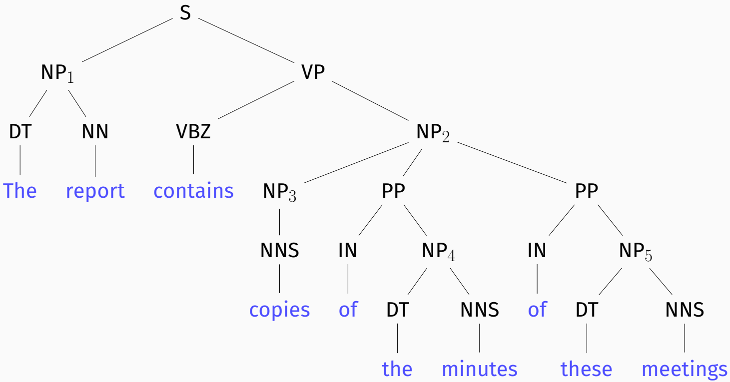
Lexical Information



Example Phrase

The report contains copies of the minutes of these meetings

Constituency Information

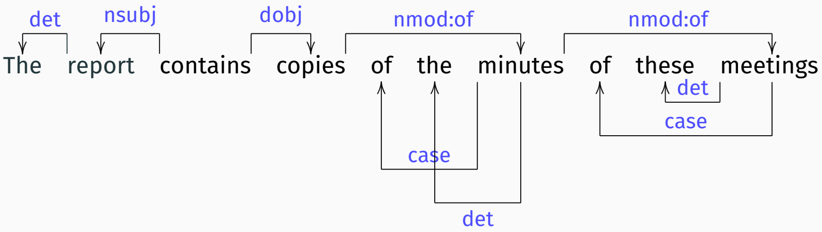


Introduction

Example Phrase

The report contains copies of the minutes of these meetings

Dependency Information



Example Phrase

The report contains copies of the minutes of these meetings

w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10}

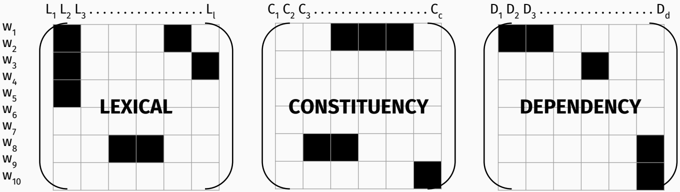
Different types of features, represented by sparse matrices

Example Phrase

The report contains copies of the minutes of these meetings

w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10}

Different types of features, represented by sparse matrices



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*

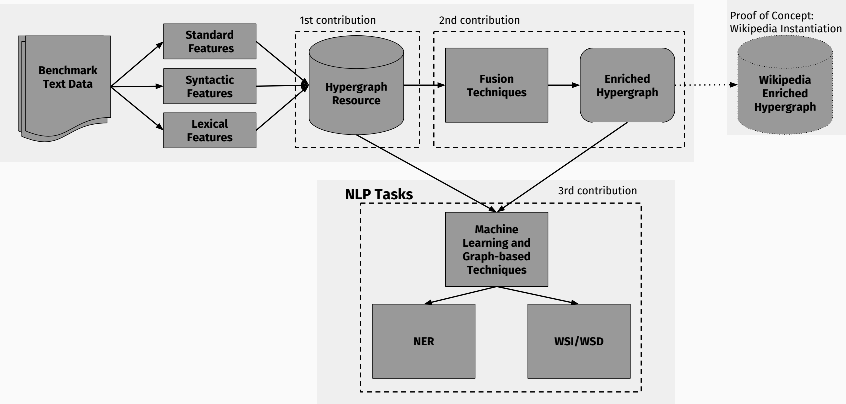
Main Challenges and Contributions

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



First Contribution: 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]

- **Generally used types of features to represent text**

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]

- **Generally used types of features to represent text**
  - Lexical

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]

- **Generally used types of features to represent text**
  - Lexical
  - Syntactic

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]

- **Generally used types of features to represent text**
  - Lexical
  - Syntactic
  - Task-specific

Example Phrase

The report contains copies of the minutes of these meetings

(These networks are going to be networks built with my example phrase)

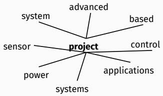
Lexical Networks

Example Phrase

The report contains copies of the minutes of these meetings

(These networks are going to be networks built with my example phrase)

Lexical Networks



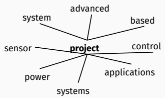
Syntactic Networks

Example Phrase

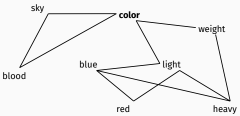
The report contains copies of the minutes of these meetings

(These networks are going to be networks built with my example phrase)

Lexical Networks



Syntactic Networks



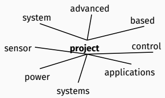
Semantic Networks

Example Phrase

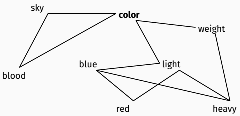
The report contains copies of the minutes of these meetings

(These networks are going to be networks built with my example phrase)

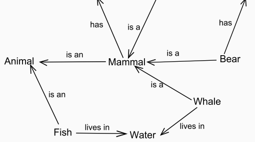
Lexical Networks



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

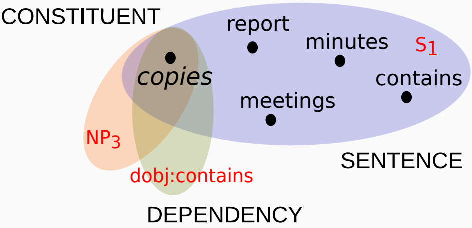
Limitations and Proposition

- **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: Hypergraph Linguistic Network



		CONSTITUENT			DEPENDENCY		SENTENCE
		NP_1 DT:NN	NP_2 NP:PP:PP	NP_3 NNS	nsubj contains	dobj contains	S_1
NN	report	1			1		1
	copies		1	1		1	1
	minutes		1				1
	meetings		1				1
VB	contains						1

Second Contribution: Combining Features and Dealing with Sparsity

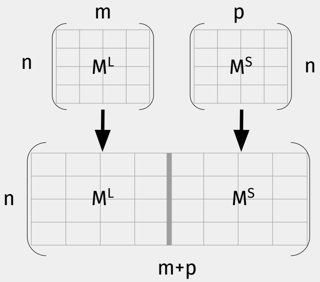
Multimedia Fusion Techniques [Atr+10; ABL10]:

- **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

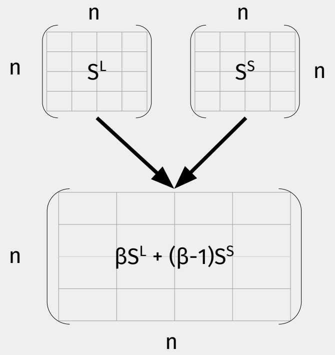
EARLY FUSION

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

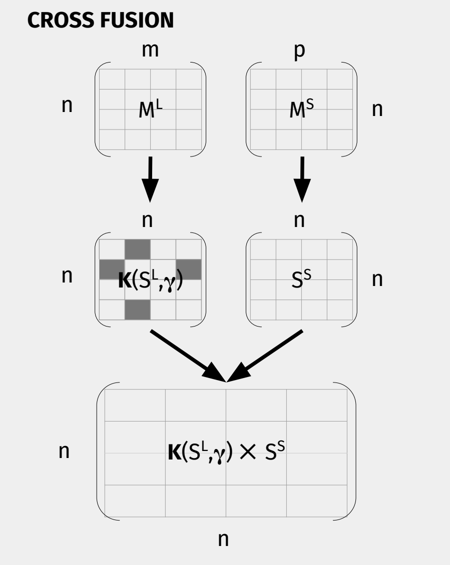


LATE FUSION: SIMILARITY FUSION

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



Cross Fusion



In our work we distinguish three levels of fusion operators:

- **First Degree Fusion (1F)**
  - $E(M^L, M^S)$
  - $X_F(S^L, M^S)$
  - $X_S(S^S, S^L)$

Levels of Fusion

In our work we distinguish three levels of fusion operators:

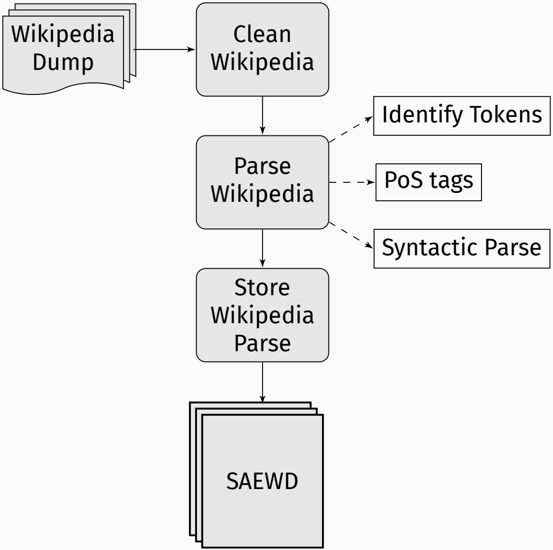
- **Second Degree Fusion (1F)**
  - Cross Feature Early Fusion: $X_F(S^L, E(M^L, M^S))$
  - Cross Feature Cross Similarity Fusion: $X_F(X_S(S^T, S^S), M^T)$
  - Early Cross Feature Fusion: $E(M^T, X_F(S^L, M^T))$
  - Late Cross Feature Fusion: $L(M^T, X_F(S^T, M^T))$

Levels of Fusion

In our work we distinguish three levels of fusion operators:

- **Higher Degree Fusion (HF)**
  - $E(M^L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M_L, X_F(S^S, M^L))))$
    - **Show decomposition of operator graphically**

SAEWD: A Wikipedia Enriched Hypergraph



Wikipedia Feature Enriched Spaces

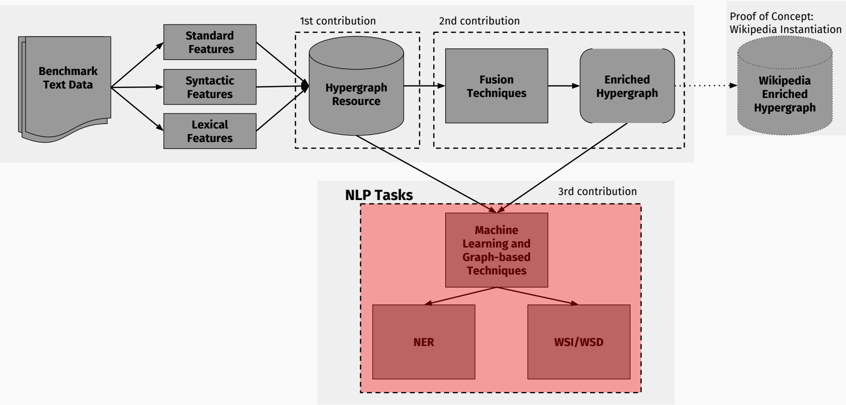
	Lexical Features (5.49%) M^L	**Syntactic Features (4.97%)** M^S	**Early Fusion (5.23%)** $E(M^L, M^S)$	X_F **Fusion (16.75%)** $X_F(S^S, M^L)$	X_F **Fusion (13.45%)** $X_F(S^L, M^S)$
	priests	monk	sailor	vassal	sailor
	nun	regent	regent	regent	fluent
priest	canton	aedile	nuclei	nun	dean
	sailor	seer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

Wikipedia Similarity Enriched Spaces

	Lexical Similarity (75.25%) S^L	**Syntactic Similarity (60.64%)** S^S	**Early Fusion (67.94%)** $E(S^L, S^S)$	**Late Fusion (83.17%)** $L(S^L, S^S)$	X_S **Fusion (87.22%)** $X_S(S^S, S^L)$	X_S **Fusion (79.69%)** $X_S(S^L, S^S)$
	wholly	regent	regent	regent	regent	sailor
	burial	coach	slang	slang	vassal	nuclei
priest	monk	broker	broker	seer	vizier	nun
	lingua	dream	rebel	tutor	leader	canton
	nuclei	tailor	tiger	cradle	result	burial

Third Contribution: Applications to Named Entity Recognition and Word Sense Disambiguation

Work Overview



Applications

- 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

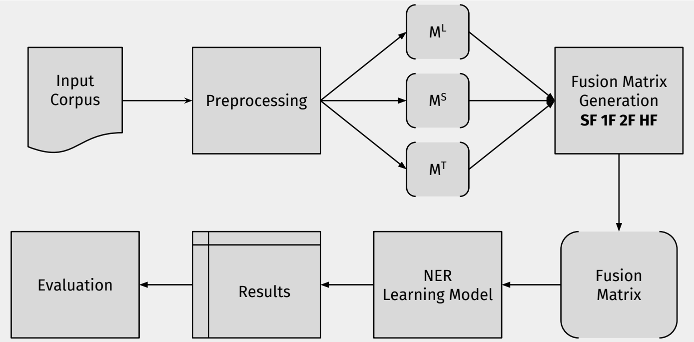
Third Contribution: Applications to Named Entity Recognition and Word Sense Disambiguation

Named Entity Recognition (NER)

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**

Experiment Flow Diagram



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

F-measure on the three datasets using single features independently with the structured perceptron

A	Single Features		
	CONLL	**WNER**	**WGLD**
M^T	**77.41**	**77.50**	**59.66**
M^L	69.40	69.17	52.34
M^S	32.95	28.47	25.49

F-measure on the three datasets using First Degree (1F) fusion operators

A	B	CONLL	WNER	WGLD
		Early Fusion (EF)		
M^L	M^S	72.01	70.59	59.38
M^L	M^T	78.13	79.78	61.96
M^S	M^T	77.70	78.10	60.93
M^L	$E(M^S, M^T)$	**78.90**	**80.04**	**63.20**
		Late Fusion (LF)		
S^L	S^S	**61.65**	58.79	44.29
S^L	S^T	55.64	**67.70**	48.00
S^S	S^T	50.21	58.41	**49.81**

		CONLL	WNER	WGLD
		Cross Feature Fusion ($X_F F$)		
S^L	M^T	49.90	**70.27**	**62.69**
S^S	M^T	47.27	51.38	48.53
S^T	$b^*_{X_F F}$	**52.89**	62.21	50.15
		Cross Similarity Fusion ($X_S F$)		
S^L	S^T	27.75	**59.12**	38.35
S^S	$b^*_{X_S F}$	36.87	40.92	39.62
S^T	$b^*_{X_S F}$	**41.89**	52.03	**39.92**

F-measure on the three datasets using Second Degree (2F) fusion operators

In $X_F X_S F$, $\hat{a}$ corresponds to the best performing matrix in the set $\{X_S(S^T, S^L), X_S(S^L, S^T), X_S(S^T, S^S)\}$

In $EX_F F$, $b^*_{EX_F F} \in \{X_F(S^S, M^L), X_F(S^L, M^L), X_F(S^L, M^T), X_F(S^S, M^L), X_F(S^S, M^T)\}$

A	B	Cross Feature Cross Similarity Fusion ($X_F X_S F$)		
		CONLL	WNER	WGLD
$\hat{a}$	M^T	37.69	59.44	**41.71**
$\hat{a}$	M^L	**38.31**	**58.73**	41.56
$\hat{a}$	M^S	29.31	52.06	34.91
		Cross Feature Early Fusion ($X_F EF$)		
		CONLL	WNER	WGLD
S^T	$E(M^L, M^T)$	**54.34**	**64.20**	39.59
S^L	$E(M^L, M^T)$	49.71	71.84	**45.14**
S^S	$E(M^L, M^T)$	47.54	53.77	43.32

		Early Cross Feature Fusion ($EX_F F$)		
		CONLL	WNER	WGLD
M^T	$b^*_{EX_F F}$	49.58	**77.32**	**61.69**
M^L	$b^*_{EX_F F}$	49.79	66.22	53.54
M^S	$b^*_{EX_F F}$	**51.53**	70.94	53.70
		Late Cross Feature Fusion ($LX_F F$)		
		CONLL	WNER	WGLD
M^T	$\hat{b}_{LX_F F}$	54.82	**75.70**	**54.73**
M^L	$\hat{b}_{LX_F F}$	**56.53**	62.27	52.39

F-measure on the three datasets using Higher Degree (HF) fusion operators

In $EEELX_FLX_F$, $\hat{b}_{EEELX_FLX_F} \in E(E(M^T, L(M^L, X_F(S^S, M^L))),$ $L(M^L, X_F(S^T, M^L))), E(E(M^T, L(M^T, X_F(S^S, M^T))), L(M^L, X_F(S^S, M^L)))$ for CONLL, WNER and WGLD.

A	B	Early Late Cross Feature Fusion (ELX_FF)		
		CONLL	WNER	WGLD
M^T	$L(M^L, X_F(S^S, M^L))$	67.16	79.45	62.37
		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

Third Contribution: Applications to Named Entity Recognition and Word Sense Disambiguation

Fusion Analysis

Analyzing the Best Fusion Operator

Decompose best fusion in four models:

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

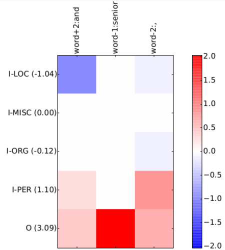
① M^{L} used to train model M_1.

② $E(\alpha_1 M^{L}, \alpha_2 M^{T})$ used to train model M_2, with $\alpha_1 = 0.95, \alpha_2 = 0.05$

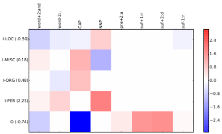
③ $E_\alpha(\alpha_1 M^{L}, \alpha_2 M^{T}, \alpha_3 L(M^{T}, X_F(S^{S}, M^{T})))$ used to train model M_3, with $\alpha_1 = 0.95, \alpha_2 = \alpha_3 = 0.05$

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 $\alpha_1 = 0.95, \alpha_2 = \alpha_3 = \alpha_4 = 0.05$

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

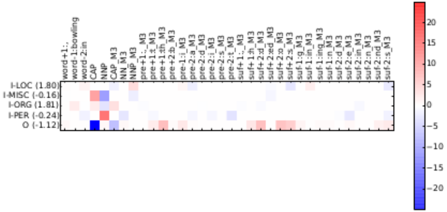


(a) The word *Kory* is predicted to be class O by model M_1 given the coefficients of the features indicated.

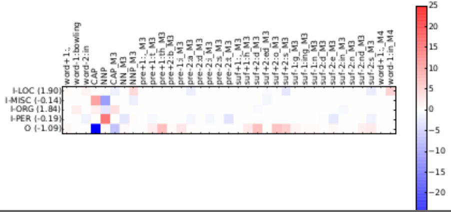


(b) The same word is correctly predicted to be class PER by model M_2 given the new features in the model.

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



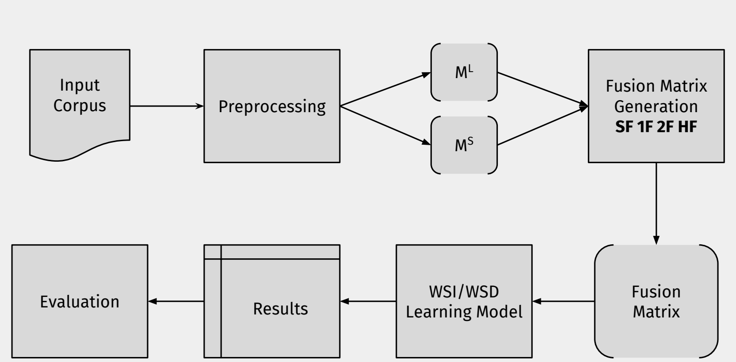
(a) The word *Green* is predicted to be class ORG by model M_3 given the coefficients of the features indicated.



Third Contribution: Applications to Named Entity Recognition and Word Sense Disambiguation

Word Sense Disambiguation

Experiment Flow Diagram



Experimental Protocol

Supevised Evaluation

Unsupevised Evaluation

Proposed Evaluation

Third Contribution: Applications to Named Entity Recognition and Word Sense Disambiguation

Leveraging the Linguistic Network Structure

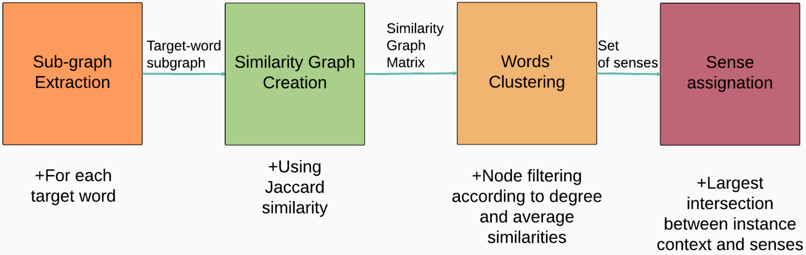
How to exploit a linguistic network to solve word sense induction and disambiguation?

- **Existing graph-based approaches**
  - Hyperlex [V04]
  - University of York (UoY) [KM07]
- **Limitations of existing approaches**
  - Single typed networks
  - Large number of parameters

Proposed Method

- **Features**
  - Automatically group words to induce senses and then assign them
  - 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
  - Use a robust and interpretable similarity measure

Proposed Method



Proposed Method: Step One

- **Creation of the linguistic network**
  - After preprocessing, we build a HLM G_{tw} that contains the co-occurrent (lexically and syntactically) words for a target word *tw*.

Proposed Method: Step Two

- **Computing the similarity between nodes**
  - G_{tw} is represented as a bipartite graph B_{tw}. Left nodes U represent words and right nodes W correspond to the hyperedges. An edge from a node u to a node w depicts the incidence of node u in hyperedge w.
  - A similarity matrix S_{tw} of dimension $|U| \times |U|$ is calculated using the Jaccard similarity: given $n_{i,j} \in U$, then $Jaccard(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$.
  - Induce a new incidence matrix F_{tw} from S_{tw} containing only the closest neighbours to each word $n_i \in U$. Each of these hyperedges represent a set of words that are deemed similar between them according to their Jaccard index value, which must be equal or higher than an assigned threshold th_1.

Proposed Method: Step Three

- **Clustering words together**
  - We select the top c-nodes in F_{tw} according to their degree. These nodes are candidate hubs, which must surpass a second threshold th_2 to be considered as proper hubs. We use the average Jaccard measure defined for each node n as:

$$AvgJaccard(n) = \frac{1}{|hedges(n)|} \sum_{h \in hedges(n)} \frac{\sum_{\substack{i \in h \\ j \in h; i \neq j}} Jaccard(i,j)}{|h+1|}$$

    where $hedeges(n)$ is the set of hyperedges n is incident in and its cardinality is defined as $|hedges(n)|$. $|h|$ is the number of nodes in hyperedge h.
  - Accepted hubs represent senses alongside with their co-occurrent words. The final set of senses is called SoS_{tw}.

Proposed Method: Step Four

- **Word Sense Disambiguation**
  - The assignation of a sense consists in looking at each *tw* instance represented by a context *ct* and simply determining which sense *s* in SoS_{tw} shares the highest amount of words with *ct*. The sense *s* is thus assigned to that instance.

Semeval Results

Unsupervised paired F-Score (FS) for Semeval-2007

FS (%)	all	nouns	verbs	#cl
1c1word	78.9	80.7	76.8	1.00
UBC-AS	78.7	80.8	76.3	1.32
DEP	74.9	80.2	69.0	3.27
LEX	61.4	62.6	60.1	4.26
UoY(2007)	56.1	65.8	45.1	9.28
Random	37.9	38.1	37.7	19.7
1c1instance	9.5	6.6	12.7	48.51

Semeval Results

Supervised Recall (SR) for Semeval-2007

SR (%)	all	nouns	verbs	#cl
I2R	81.6	86.8	75.7	3.08
LEX	79.4	82.5	75.9	4.26
DEP	79.1	81.5	76.4	3.27
MFS	78.7	80.9	76.2	1
UoY(2007)	77.7	81.6	73.3	9.28

Semeval Results

- **Discussion**
  - Both **DEP** and **LEX** beat the competition baselines
  - They also beat the most similar approach UoY(2007)
  - Best result for verbs concerning supervised Recall
  - Possibility for features' combination: both seem to complement each other

Conclusions and Future Work

Conclusions

Future Work

References

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

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.

References II

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.

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 III

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 IV

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.

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.

References V

Carina Silberer and Simone Paolo Ponzetto. "UHD: Cross-lingual Word Sense Disambiguation Using Multilingual Co-occurrence Graphs". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 134–137. URL: http://dl.acm.org/citation.cfm?id=1859664.1859691.

Antonio Di Marco and Roberto Navigli. "Clustering Web Search Results with Maximum Spanning Trees". In: *Proceedings of the 12th International Conference on Artificial Intelligence Around Man and Beyond*. AI*IA'11. Palermo, Italy: Springer-Verlag, 2011, pp. 201–212. ISBN: 978-3-642-23953-3. URL: http://dl.acm.org/citation.cfm?id=2041977.2042002.

References VI

David Jurgens. "Word Sense Induction by Community Detection". In: *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*. TextGraphs-6. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 24–28. ISBN: 978-1-937284-008. URL: http://dl.acm.org/citation.cfm?id=2024277.2024282.

Antoon Bronselaer and Gabriella Pasi. "An approach to graph-based analysis of textual documents". In: *Proceedings of the 8th conference of the European Society for Fuzzy Logic and Technology, EUSFLAT-13, Milano, Italy, September 11-13, 2013*. 2013. DOI: 10.2991/eusflat.2013.96.

Avneesh Saluja and Jirı Navrátil. "Graph-Based Unsupervised Learning of Word Similarities Using Heterogeneous Feature Types". In: *Graph-Based Methods for Natural Language Processing* (2013), p. 29.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. "Entity Linking meets Word Sense Disambiguation: a Unified Approach". In: *Transactions of the Association for Computational Linguistics (TACL)* 2 (2014), pp. 231–244.

Tao Qian et al. "Word Sense Induction Using Lexical Chain based Hypergraph Model". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014, pp. 1601–1611. URL: http://www.aclweb.org/anthology/C14-1152.

Appendix

Appendix

WSI/D Method in Detail

Proposed Method: Step One

- **Creation of the linguistic network**
  - After preprocessing, we build a HLM G_{tw} that contains the co-occurrent (lexically and syntactically) words for a target word tw.

- **Computing the similarity between nodes**
  - G_{tw} is represented as a bipartite graph B_{tw}. Left nodes U represent words and right nodes W correspond to the hyperedges. An edge from a node u to a node w depicts the incidence of node u in hyperedge w.
  - A similarity matrix S_{tw} of dimension $|U| \times |U|$ is calculated using the Jaccard similarity: given $n_{i,j} \in U$, then $Jaccard(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$.
  - Induce a new incidence matrix F_{tw} from S_{tw} containing only the closest neighbours to each word $n_i \in U$. Each of these hyperedges represent a set of words that are deemed similar between them according to their Jaccard index value, which must be equal or higher than an assigned threshold th_1.

Proposed Method: Step Three

- **Clustering words together**
  - We select the top c-nodes in F_{tw} according to their degree. These nodes are candidate hubs, which must surpass a second threshold th_2 to be considered as proper hubs. We use the average Jaccard measure defined for each node n as:

$$AvgJaccard(n) = \frac{1}{|hedges(n)|} \sum_{h \in hedges(n)} \frac{\sum_{\substack{i \in h \\ j \in h; i \neq j}} Jaccard(i,j)}{|h+1|}$$

    where $hedeges(n)$ is the set of hyperedges n is incident in and its cardinality is defined as $|hedges(n)|$. $|h|$ is the number of nodes in hyperedge h.
  - Accepted hubs represent senses alongside with their co-occurrent words. The final set of senses is called SoS_{tw}.

Proposed Method: Step Four

- **Word Sense Disambiguation**
  - The assignation of a sense consists in looking at each *tw* instance represented by a context *ct* and simply determining which sense *s* in SoS_{tw} shares the highest amount of words with *ct*. The sense *s* is thus assigned to that instance.

Experiments

- **Implementation Framework**
  - **Systems built and evaluated**: **DEP** and **LEX**.
    - **DEP**: Syntactical dependencies
    - **LEX**: Lexical co-occurrences
  - **Two datasets**: Semeval-2007 Task 2 (100 words: 35 nouns, 65 verbs) and Semeval-2010 Task 14 (100 words: 50 nouns, 50 verbs). For brevity, only the results for the first dataset are discussed in this presentation.
  - **Evaluation metrics**: Unsupervised evaluation (Paired F-Score, V-Measure). Supervised evaluation (Recall).

Results Semeval 2010

VM (%)	all	nouns	verbs	#cl
Hermit	16.2	16.7	15.6	10.78
NMF_{lib}	11.8	13.5	9.4	4.80
LEX	11.6	8.8	11.9	10.5
Random	4.4	4.2	4.6	4.00
DEP	3.5	3.9	2.8	2.75
MFS	0.0	0.0	0.0	1.00

Table 1: Unsupervised V-Measure (VM) on the Semeval 2010 test set

Results Semeval 2010

FS (%)	all	nouns	verbs	#cl
MFS	63.5	57.0	72.4	1.00
Duluth-WSI-SVD-Gap	63.3	57.0	72.4	1.02
DEP	53.6	50.1	58.7	2.75
NMF_{lib}	45.3	42.2	49.8	5.42
LEX	38.4	46.7	28.5	10.5
Random	31.9	30.4	34.1	4.00

Table 2: Unsupervised Paired F-Score (FS) for the Semeval 2010 test set

Results Semeval 2010

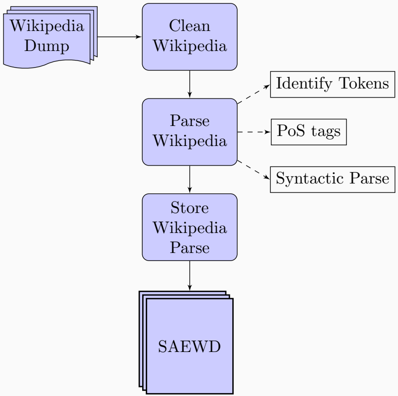
SR (%)	all	nouns	verbs
NMF$_{lib}$	62.6	57.3	70.2
UoY(2010)	62.4	59.4	66.8
LEX	59.8	55.8	67.4
DEP	59.3	53.9	67.2
MFS	58.7	53.2	66.6
Random	57.3	51.5	65.7

Table 3: Supervised recall (SR) for Semeval 2010 test set (80% mapping, 20% evaluation)

Appendix

SAEWD

Building SAEWD



SAEWD: Parsed sample

FILENAME wiki_00.parsed

token	lemma	POS	constituency	head	dependency
%%#PAGE Anarchism					
⋮	⋮	⋮	⋮	⋮	⋮
%%#SEN 25 9					
A	a	DT	NP_22,S_97	3	det
great	great	JJ	NP_22,S_97	3	amod
brigand	brigand	NN	NP_22,S_97	4	nsubj
becomes	become	VBZ	VP_44,S_97	0	root
a	a	DT	NP_18,NP_20,VP_44,S_97	6	det
ruler	ruler	NN	NP_18,NP_20,VP_44,S_97	4	xcomp
of	of	IN	PP_57,NP_20,VP_44,S_97	9	case
a	a	DT	NP_18,PP_57,NP_20,VP_44,S_97	9	det
Nation	nation	NN	NP_18,PP_57,NP_20,VP_44,S_97	6	nmod

Appendix

Ongoing Results

Ongoing Work: Results

- **Combining the hyperedges: cross fusion**

  Unsupervised paired F-Score (FS) for the Semeval 2007 test set

FS (%)	**all**	**nouns**	**verbs**	**#cl**
1c1word	78.9	80.7	76.8	1.00
UBC-AS	78.7	80.8	76.3	1.32
CROSS$_{k=75}$	78.6	80.7	76.3	1.70
DEP	74.9	80.2	69.0	3.27
CLUST$_{k=5,th=55}$	72.5	76.0	63.8	5.47
LEX	61.4	62.6	60.1	4.26
UoY(2007)	56.1	65.8	45.1	9.28
Random	37.9	38.1	37.7	19.7
1c1instance	9.5	6.6	12.7	48.51