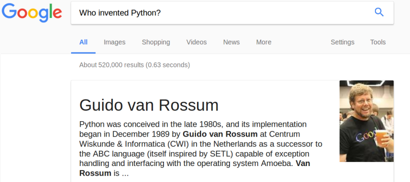
Hypergraphs and Information Fusion for Term Representation Enrichment. Applications to Named Entity Recognition and Word Sense Disambiguation

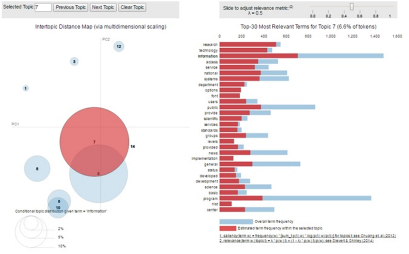
Ph.D. Thesis Defense

Pavel Soriano-Morales
Supervised by Sabine Loudcher and Julien Ah-Pine

February 7th, 2018

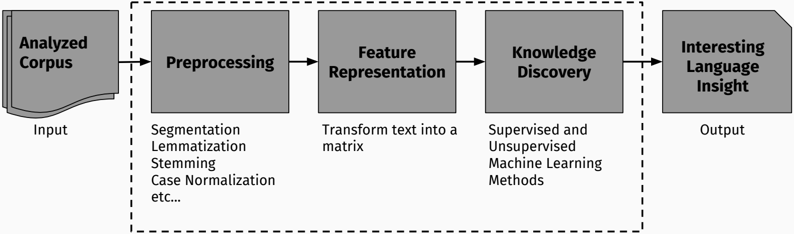
Why is it useful to us to understand text?



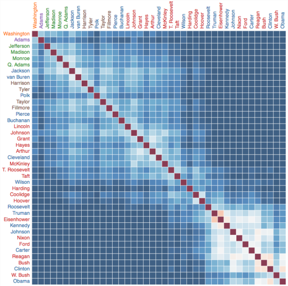


How do we extract meaning from text?

We use **Natural Language Processing** (NLP), a field of computer science interested in making computers comprehend text and obtain useful information from it



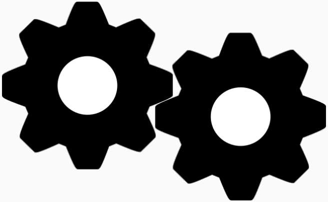
How do we represent text for the machine to understand?



Dealing with data sparsity

Leveraging heterogeneity

What techniques do we use to discover meaning from text?



Finding semantic communities

Representing Text

- **Common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
  - Semantic

Representing Text

- **Common ways to represent text**
    - Lexical
    - Syntactic
        - Constituency Tree
        - Dependency Tree
    - Semantic
- **Example Phrase**

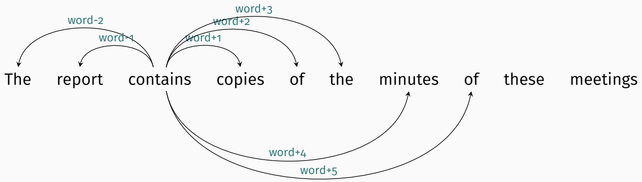
    The report contains copies of the minutes of these meetings

Representing Text

- **Common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
  - Semantic
- **Example Phrase**

  The report contains copies of the minutes of these meetings

Lexical Representation

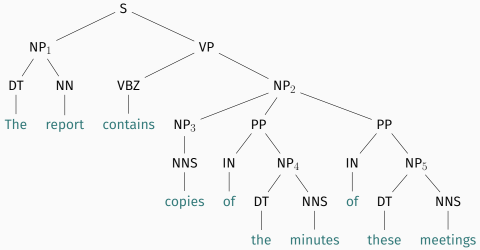


Representing Text

- **Common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
  - Semantic
- **Example Phrase**

  The report contains copies of the minutes of these meetings

Constituency Tree Representation

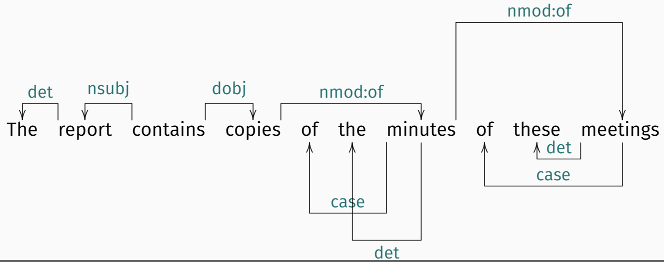


Representing Text

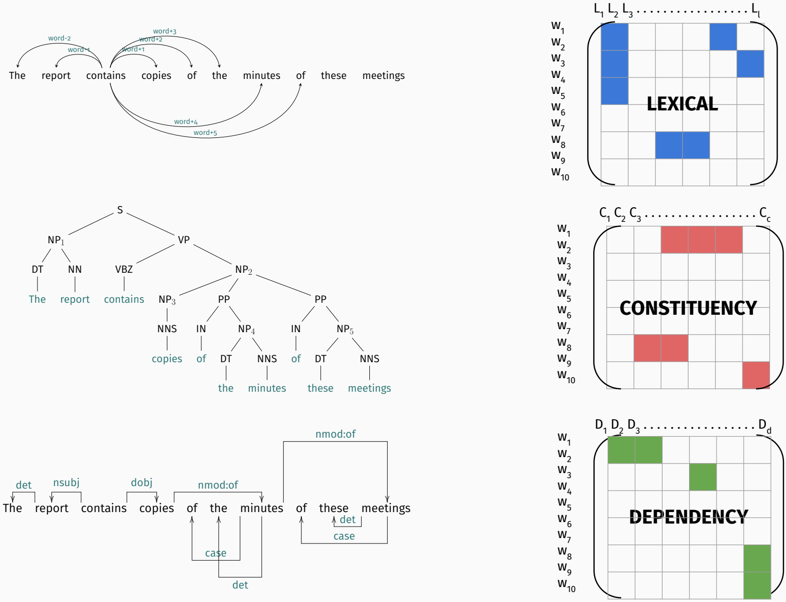
- **Common ways to represent text**
  - Lexical
  - Syntactic
    - Constituency Tree
    - Dependency Tree
  - Semantic
- **Example Phrase**

  The report contains copies of the minutes of these meetings

Dependency Tree Representation



Represention Models



Main Challenges and Contributions

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

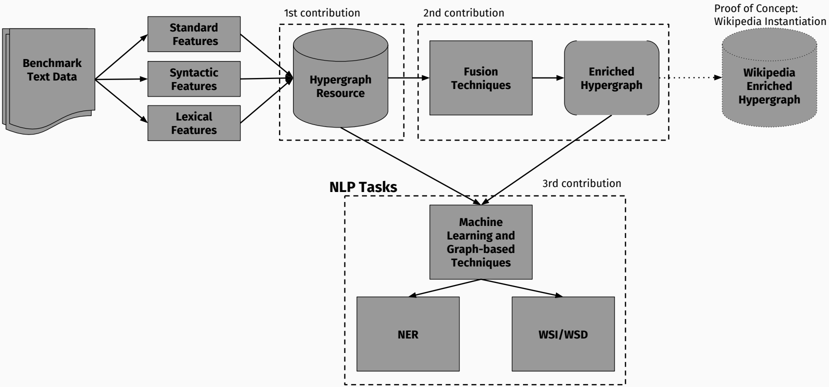
Main Challenges and Contributions

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

- **Leveraging contexts**
    - We extract linguistic information from words based on the distributional hypothesis (a word is defined by its surroundings)
    - These surroundings are defined as contexts
    - Contexts are formed by the interactions a word participates in. These interactions can be lexical or syntactical or other types.

Introduction

- **Leveraging contexts**
  - We extract linguistic information from words based on the distributional hypothesis (a word is defined by its surroundings)
  - These surroundings are defined as contexts
  - Contexts are formed by the interactions a word participates in. These interactions can be lexical or syntactical or other types.
- **We use network models to represent contexts**
  - Graphs structures can give us a clearer view into the relations of words within a text
  - Allow us to apply methods from graph theory
  - Ultimately graphs are transformed to a vectorial representation through the adjacency/incidence matrices

Example phrase

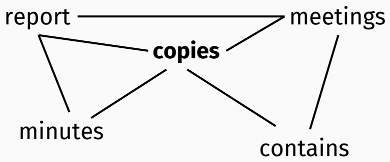
The report contains copies of the minutes of these meetings

Example phrase

The report contains copies of the minutes of these meetings

Lexical Networks

Sentence Level

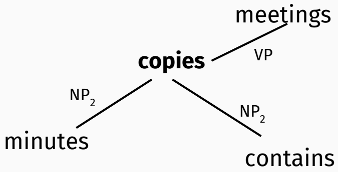


Example phrase

The report contains copies of the minutes of these meetings

Syntactic Networks

Constituency Tree

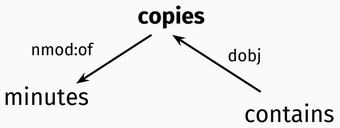


Example phrase

The report contains copies of the minutes of these meetings

Syntactic Networks

Dependency Tree



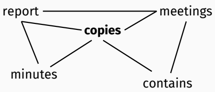
- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network relate maximum two words at each time

Limitations and Proposition

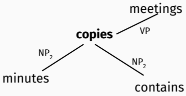
- **Limitations of existing representations**
  - Language networks generally employ a single type of textual information
  - The edges of the network relate maximum two words at each time
- **Proposition**
  - Use a hypergraph model to link together the different types of networks
  - This allows for a semantic overview at three different layers: short range, medium range, and long range at once
  - Relating more than two words at the same time

Proposed Model

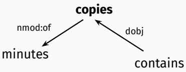
Lexical Networks
Sentence Level



Syntactic Networks
Constituency Tree

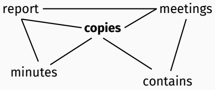


Syntactic Networks
Dependency Tree

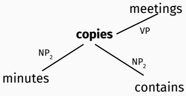


Proposed Model

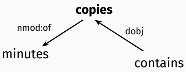
Lexical Networks
Sentence Level



Syntactic Networks
Constituency Tree



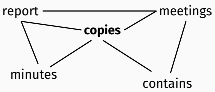
Syntactic Networks
Dependency Tree



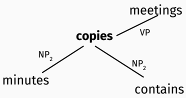
Hypergraph Model

Proposed Model

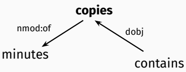
Lexical Networks
Sentence Level



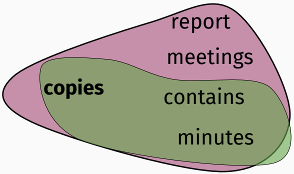
Syntactic Networks
Constituency Tree



Syntactic Networks
Dependency Tree



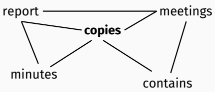
Hypergraph Model



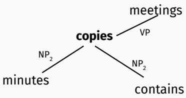
- Lexical
- Constituency (NP_2)

Proposed Model

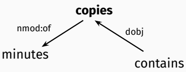
Lexical Networks
Sentence Level



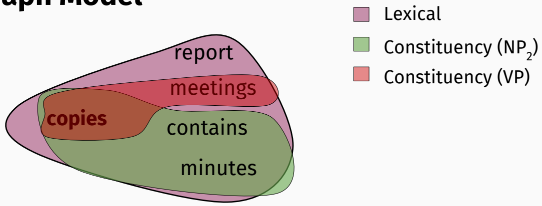
Syntactic Networks
Constituency Tree



Syntactic Networks
Dependency Tree

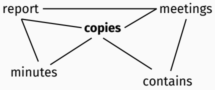


Hypergraph Model

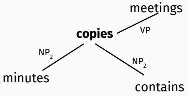


Proposed Model

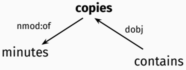
Lexical Networks
Sentence Level



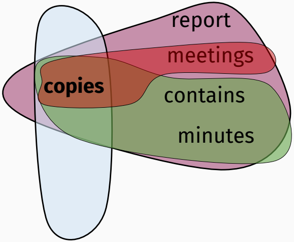
Syntactic Networks
Constituency Tree



Syntactic Networks
Dependency Tree



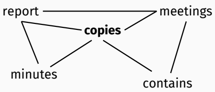
Hypergraph Model



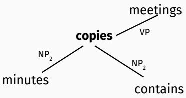
- Lexical
- Constituency (NP_2)
- Constituency (VP)
- Dependency (dobj:contains)

Proposed Model

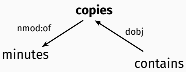
Lexical Networks
Sentence Level



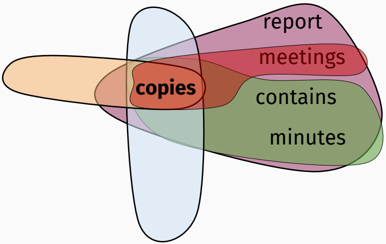
Syntactic Networks
Constituency Tree



Syntactic Networks
Dependency Tree



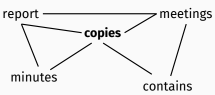
Hypergraph Model



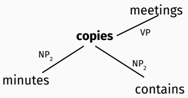
- Lexical
- Constituency (NP_2)
- Constituency (VP)
- Dependency (dobj:contains)
- Dependency (nmod:of)

Proposed Model

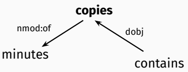
Lexical Networks
Sentence Level



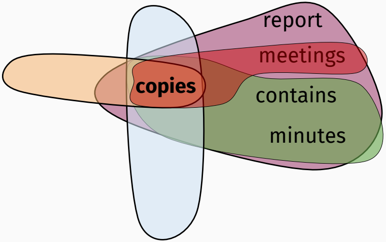
Syntactic Networks
Constituency Tree



Syntactic Networks
Dependency Tree



Hypergraph Model



- Lexical
- Constituency (NP_2)
- Constituency (VP)
- Dependency (dobj:contains)
- Dependency (nmod:of)

How to combine these heterogeneous networks into a single representation?

Contributions in Detail

Combining Features and Dealing with Sparsity

- **Definition**
  - Used in multimedia analysis tasks to integrate multiple media
  - We adapt them to combine textual information
  - The goal is to obtain rich insights about the data being treated
  - By creating a single representation from heterogeneous information

Multimedia Fusion Techniques

- **Definition**
  - Used in multimedia analysis tasks to integrate multiple media
  - We adapt them to combine textual information
  - The goal is to obtain rich insights about the data being treated
  - By creating a single representation from heterogeneous information
- **Main fusion operators:**
  - Early Fusion $E_\alpha(\cdot)$,
  - Late Fusion $L_\beta(\cdot)$,
  - Cross Fusion $X_\gamma(\cdot)$

Early and Late Fusion

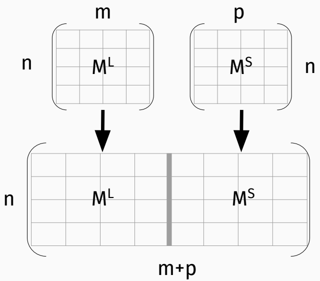
DEFINITIONS

M^L	Lexical features
S^L	Lexical similarities

M^S	Syntactic features
S^S	Syntactic similarities

EARLY FUSION

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



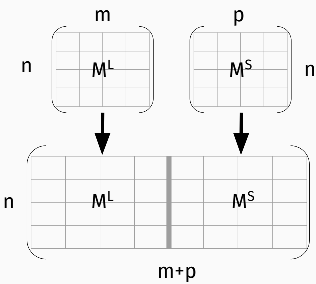
Early and Late Fusion

DEFINITIONS

M^L	Lexical features	M^S	Syntactic features
S^L	Lexical similarities	S^S	Syntactic similarities

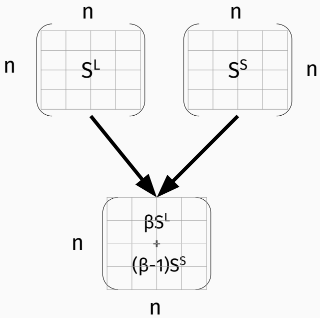
EARLY FUSION

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

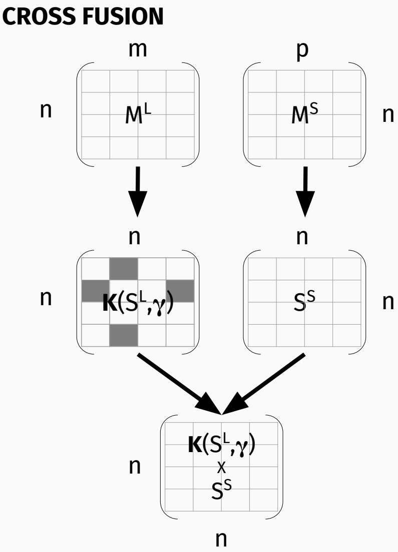


LATE FUSION: SIMILARITY FUSION

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



Cross Fusion



Hybrid Fusion

- **Combining fusion operators**
    - Applying one function to the result of another to produce a new fusion function

Hybrid Fusion

- **Combining fusion operators**
  - Applying one function to the result of another to produce a new fusion function
- **First Degree**
  - $E(M^L, M^S)$, $L(S^S, M^L)$
  - **Cross Feature Fusion**: $X_F(S^S, M^L)$
  - **Cross Similarity Fusion**: $X_S(S^S, S^L)$

$X_F(S^L, M^S)$ Cross Feature Fusion

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array}\begin{array}{c} w_1\ w_2\ w_3 \\ \left(\begin{array}{c} \\ S^L \\ \\ \end{array}\right) \end{array} \times \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array}\begin{array}{c} f_{S1}\ f_{S2}\ f_{S3} \\ \left(\begin{array}{c} \\ M^S \\ \\ \end{array}\right) \end{array} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array}\begin{array}{c} f_{S1}\ f_{S2}\ f_{S3} \\ \left(\begin{array}{c} \\ X_F(S^L, M^S) \\ \\ \end{array}\right) \end{array}$$

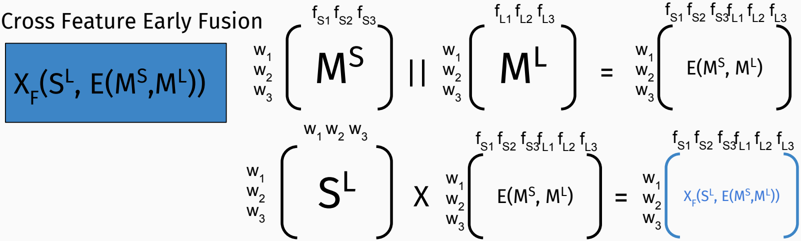
$X_S(S^L, S^S)$ Cross Similarity Fusion

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array}\begin{array}{c} w_1\ w_2\ w_3 \\ \left(\begin{array}{c} \\ S^L \\ \\ \end{array}\right) \end{array} \times \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array}\begin{array}{c} w_1\ w_2\ w_3 \\ \left(\begin{array}{c} \\ S^S \\ \\ \end{array}\right) \end{array} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array}\begin{array}{c} w_1\ w_2\ w_3 \\ \left(\begin{array}{c} \\ X_S(S^L, S^S) \\ \\ \end{array}\right) \end{array}$$

Hybrid Fusion

- **Combining fusion operators**
  - Applying one function to the result of another to produce a new fusion function

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



- **Combining fusion operators**
  - Applying one function to the result of another to produce a new fusion function

- **Higher Degree**
  - Triple Early Double Late Cross Feature Fusion: $E(M_L, E(E(M_T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$

Higher Degree Operator

$$E(M_L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$$

Higher Degree Operator

$$E(M_L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$$

Higher Degree Operator

$$E(M_L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$$

$$L(M^L, X_F(S^S, M^L))$$

$$\begin{array}{cc} & \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} \quad S^S \quad \end{bmatrix} \end{array} \times \begin{array}{cc} & \begin{array}{ccc} f_{L1} & f_{L2} & f_{L3} \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} \quad M^L \quad \end{bmatrix} \end{array} = \begin{array}{cc} & \begin{array}{ccc} f_{L1} & f_{L2} & f_{L3} \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} X_F(S^S, M^L) \end{bmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccc} f_{L1} & f_{L2} & f_{L3} \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} \quad M^L \quad \end{bmatrix} \end{array} + \begin{array}{cc} & \begin{array}{ccc} f_{L1} & f_{L2} & f_{L3} \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} X_F(S^S, M^L) \end{bmatrix} \end{array} = \begin{array}{cc} & \begin{array}{ccc} f_{L1} & f_{L2} & f_{L3} \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} L(M^L, X_F(S^S, M^L)) \end{bmatrix} \end{array}$$

Higher Degree Operator

$$E(M_L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$$

$$L(M^T, X_F(S^T, M^T)))$$

$$\begin{array}{c|c} & w_1\ w_2\ w_3 \\ \hline \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \begin{pmatrix} S^T \end{pmatrix} \end{array} \times \begin{array}{c|c} & f_{T1}\ f_{T2}\ f_{T3} \\ \hline \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \begin{pmatrix} M^T \end{pmatrix} \end{array} = \begin{array}{c|c} & f_{T1}\ f_{T2}\ f_{T3} \\ \hline \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \begin{pmatrix} X_F(S^T, M^T) \end{pmatrix} \end{array}$$

$$\begin{array}{c|c} & f_{T1}\ f_{T2}\ f_{T3} \\ \hline \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \begin{pmatrix} M^T \end{pmatrix} \end{array} + \begin{array}{c|c} & f_{T1}\ f_{T2}\ f_{T3} \\ \hline \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \begin{pmatrix} X_F(S^T, M^T) \end{pmatrix} \end{array} = \begin{array}{c|c} & f_{T1}\ f_{T2}\ f_{T3} \\ \hline \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \begin{pmatrix} L(M^T, X_F(S^T, M^T)) \end{pmatrix} \end{array}$$

Higher Degree Operator

$$\boxed{E(M_L, \boxed{E(\boxed{E(M^T, \boxed{L(M^T, X_F(S^T, M^T))})}, \boxed{L(M^L, X_F(S^S, M^L))})})}$$

$$\boxed{E(M^T, L(M^T, X_F(S^T, M^T)))}$$

$$\begin{array}{c c} & f_{T1}\, f_{T2}\, f_{T3} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \left[\; M^T \;\right] \end{array} \;||\; \begin{array}{c c} & f_{T1}\, f_{T2}\, f_{T3} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \left[\; L(M^T, X_F(S^T, M^T)) \;\right] \end{array} = \begin{array}{c c} & f_{T1}\, f_{T2}\, f_{T3} \quad f_{T1}\, f_{T2}\, f_{T3} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \left[\; E(M^T, L(M^T, X_F(S^T, M^T))) \;\right] \end{array}$$

Higher Degree Operator

$$\boxed{E(M_L, \boxed{E(\boxed{E(M^T, \boxed{L(M^T, X_F(S^T, M^T))})}, \boxed{L(M^L, X_F(S^S, M^L))})})}$$

$$\boxed{E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))}$$

$$\begin{array}{cc} & f_{T1}\ f_{T2}\ f_{T3} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \left(E(M^T, L(M^T, X_F(S^T, M^T))) \right) \end{array} \; || \; \begin{array}{cc} & f_{L1}\ f_{L2}\ f_{L3} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \left(L(M^L, X_F(S^S, M^L)) \right) \end{array} =$$

$$\begin{array}{cc} & f_{L1}\ f_{L2}\ f_{L3}\quad f_{L1}\ f_{L2}\ f_{L3} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \left(E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L)))) \right) \end{array}$$

Higher Degree Operator

$$E(M_L, \boxed{E(\boxed{E(M^T, \boxed{L(M^T, X_F(S^T, M^T))})}, \boxed{L(M^L, X_F(S^S, M^L))})})$$

$$E(M_L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$$

$$\begin{array}{cc} & f_{L1}\ f_{L2}\ f_{L3} \\ \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \left(M^T \right) \end{array} \; || \; \begin{array}{cc} & f_{L1}\ f_{L2}\ f_{L3} \\ \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \left(E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L)))) \right) \end{array} =$$

$$\begin{array}{cc} & f_{L1}\ f_{L2}\ f_{L3}\ f_{L1}\ f_{L2}\ f_{L3} \\ \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} & \left(E(M_L, E(E(M^T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L)))) \right) \end{array}$$

Contributions in Detail

Finding Communities in the Network

- **Language networks tend to be scale-free**
  - There are certain nodes (hubs) that are very well connected forming communities within the network

Introduction

- **Language networks tend to be scale-free**
  - There are certain nodes (hubs) that are very well connected forming communities within the network
- **Seminal approaches**
  - Hyperlex [V04]
  - University of York (UoY) [KM07]

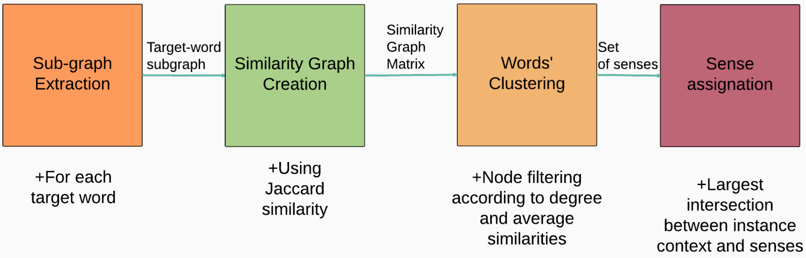
Introduction

- **Language networks tend to be scale-free**
  - There are certain nodes (hubs) that are very well connected forming communities within the network
- **Seminal approaches**
  - Hyperlex [V04]
  - University of York (UoY) [KM07]
- **Limitations of existing approaches**
  - Single typed networks
  - Large number of parameters

Introduction

- **Language networks tend to be scale-free**
  - There are certain nodes (hubs) that are very well connected forming communities within the network
- **Seminal approaches**
  - Hyperlex [V04]
  - University of York (UoY) [KM07]
- **Limitations of existing approaches**
  - Single typed networks
  - Large number of parameters
- **Proposition**
  - 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



Applications to NLP

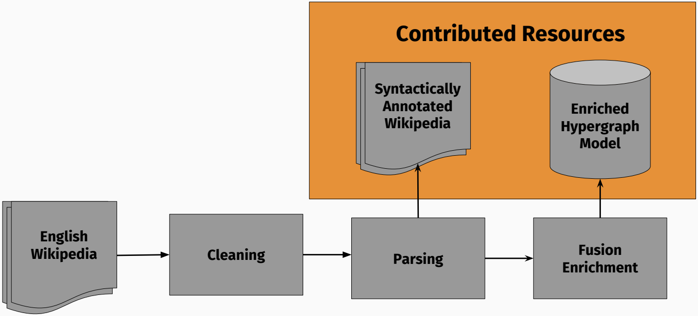
Hypergraph Model Instantiation

Hypergraph Model Instantiation

- **Apply our proposed linguistic model to a real world corpus**
  - Use the English Wikipedia as input and generate a textual structure following the proposed network model

Hypergraph Model Instantiation

- **Apply our proposed linguistic model to a real world corpus**
  - Use the English Wikipedia as input and generate a textual structure following the proposed network model
- **Steps performed**



Hypergraph Incidence Matrix

		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

Wikipedia Feature Enriched Space

- **Characteristics of the enriched space**
  - Sparsity is reduced
  - Semantic relatedness differs according to the representation space

	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

- **NER Objective**
  - The goal is to automatically discover mentions that belong to a well-defined semantic category.

Introduction

- **NER Objective**
  - The goal is to automatically discover mentions that belong to a well-defined semantic category.
- **Classic entities types**
  - Location (LOC)
  - Organization (ORG)
  - Person (PER)
  - Miscellaneous (MISC)
  - None (O)

Introduction

- **NER Objective**
  - The goal is to automatically discover mentions that belong to a well-defined semantic category.
- **Classic entities types**
  - Location (LOC)
  - Organization (ORG)
  - Person (PER)
  - Miscellaneous (MISC)
  - None (O)
- **Our goal**
  - We assess the effectiveness of the classic fusion methods and propose new hybrid combinations

Example Phrase

Australian scientist discovers star with telescope

Three different types of features

Word	Features	Feature Type
Australian	word:Australian, word+1:scientist, ...	**Lexical (L)**
scientist	Australian/JJ/amod, discovers/VBZ/nsubj_inv	**Syntactic (S)**
discover	discover, no-capital-letter, prf:dis, suf:ver, VBZ	**Standard (T)**

Experimental Protocol

- **Preprocessing**
  - Normalize numbers

Experimental Protocol

- **Preprocessing**
  - Normalize numbers
- **Test Corpora**
  - CoNLL-2003 (CONLL): Train: 219,554 lines. Test: 50,350 lines
  - Wikiner (WNER): 3.5 million words.
  - Wikigold (WGLD): 41,011 words.

Experimental Protocol

- **Preprocessing**
  - Normalize numbers
- **Test Corpora**
  - CoNLL-2003 (CONLL): Train: 219,554 lines. Test: 50,350 lines
  - Wikiner (WNER): 3.5 million words.
  - Wikigold (WGLD): 41,011 words.
- **Learning Algorithm**
  - Structured Perceptron

Experimental Protocol

- **Preprocessing**
  - Normalize numbers
- **Test Corpora**
  - CoNLL-2003 (CONLL): Train: 219,554 lines. Test: 50,350 lines
  - Wikiner (WNER): 3.5 million words.
  - Wikigold (WGLD): 41,011 words.
- **Learning Algorithm**
  - Structured Perceptron
- **Evaluation Metric**
  - F-measure
  - Evaluated with a 5-fold CV (WNER and WGLD)

Evaluation Baselines (F-measure)

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

A	*B*	Early Fusion (EF)		
		CONLL	**WNER**	**WGLD**
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

A	B	Baseline (EF)		
		CONLL	**WNER**	**WGLD**
M^L	$E(M^S, M^T)$	78.90	80.04	63.20

First Degree Fusion

		Cross Feature Fusion ($X_F F$)		
		CONLL	**WNER**	**WGLD**
S^L	M^T	49.90	**70.27**	**62.69**
S^S	M^T	47.27	51.38	48.53
S^T	M^L	**52.89**	62.21	50.15

A	B	Baseline (EF)		
		CONLL	**WNER**	**WGLD**
M^L	$E(M^S, M^T)$	78.90	80.04	63.20

Second Degree Fusion

A	B	Early Cross Feature Fusion (EX_FF)		
		CONLL	**WNER**	**WGLD**
M^T	$X_F(S^S, M^L)$	49.58	**77.32**	**61.69**

Evaluation (F-measure)

A	B	Baseline (EF)		
		CONLL	WNER	WGLD
M^L	$E(M^S, M^T)$	78.90	80.04	63.20

Second Degree Fusion

A	B	Late Cross Feature Fusion ($LX_F F$)		
		CONLL	WNER	WGLD
M^T	$X_F(S^S, M^T)$	**56.53**	62.27	52.39

Evaluation (F-measure)

A	B	Baseline (EF)		
		CONLL	WNER	WGLD
M^L	$E(M^S, M^T)$	78.90	80.04	63.20

High Degree Fusion

		Triple Early Double Late Cross Feature Fusion ($\mathbf{EEELX_FLX_F}$)		
		CONLL	WNER	WGLD
$\mathbf{M}^{\mathbf{L}}_{\alpha=\mathbf{0.95}}$	$\hat{\mathbf{b}}_{\mathbf{EEELX_FLX_F}}$	**79.67**	**81.79**	**67.05**

$$\hat{b}_{EEELX_FLX_F} = E(E(M^T, L(M^T, X_F(S^S, M^T))), L(M^L, X_F(S^S, M^L)))$$

Analyzing the Best Fusion Operator

- **Split the operator in four different models**

$$\overbrace{\underbrace{\overbrace{E_{\alpha=0.95}(\underbrace{M^L}_{M_1}, M^T}^{M_2}, L(M^T, X_F(S^S, M^T))}_{M_3}, L(M^L, X_F(S^S, M^L)))}^{M_4}$$

Analyzing the Best Fusion Operator

- **Split the operator in four different models**

$$\overbrace{\underbrace{\overbrace{E_{\alpha=0.95}(\underbrace{M^L}_{M_1}, M^T}^{M_2}, L(M^T, X_F(S^S, M^T))}_{M_3}, L(M^L, X_F(S^S, M^L)))}^{M_4}$$

M_1 M^L

M_2 $E_\alpha(M^L, M^T)$

M_3 $E_\alpha(M^L, M^T, L(M^T, X_F(S^S, M^T)))$

M_4 $E_\alpha(M^L, M^T, L(M^T, X_F(S^S, M^T)), L(M^L, X_F(S^S, M^L)))$

Analyzing the Best Fusion Operator

- **Error Analysis Model**
  - To facilitate the interpretation, we change the prediction model to a logistic regression with L_1 normalization, which also benefits from the enriched spaces

Analyzing the Best Fusion Operator

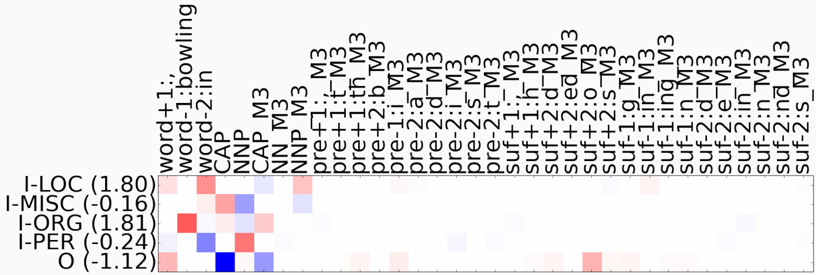
- **Error Analysis Model**
    - To facilitate the interpretation, we change the prediction model to a logistic regression with L_1 normalization, which also benefits from the enriched spaces
- **Procedure**
    - We find an error on a model and then see if this error was fixed in the next evolved model
    - We study the weights assigned to each feature and see if those added by the fusion make the model change its decision

Analyzing the Best Fusion Operator

- **Error Analysis Model**
  - To facilitate the interpretation, we change the prediction model to a logistic regression with L_1 normalization, which also benefits from the enriched spaces
- **Procedure**
  - We find an error on a model and then see if this error was fixed in the next evolved model
  - We study the weights assigned to each feature and see if those added by the fusion make the model change its decision
- **Experiment**
  - We follow the location name *Green* from M_3 (incorrectly classified as ORG) to M_4 (correctly classified as LOC)

Analyzing the Best Fusion Operator

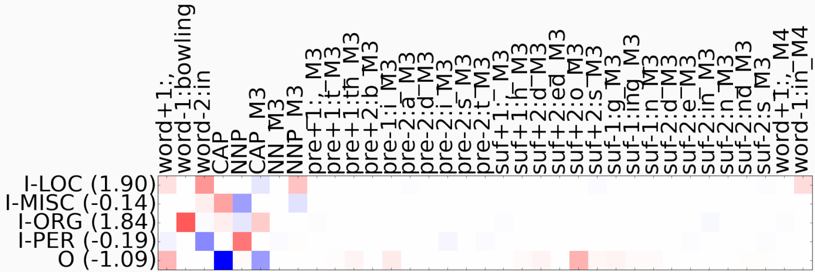
The location *Green* is classified as ORG by M_3. It is fixed by M_4, classifying it as LOC



M_3

Analyzing the Best Fusion Operator

The location *Green* is classified as ORG by M_3. It is fixed by M_4, classifying it as LOC



M_4

Applications to NLP

Solving Word Sense Induction and Disambiguation

- **WSI/WSD Objective**
    - The goal is to determine a set of possible senses to a given word according to its possible contexts (WSI). Then, assigning a correct sense to a particular instance of said word (WSD)

Introduction

- **WSI/WSD Objective**
  - The goal is to determine a set of possible senses to a given word according to its possible contexts (WSI). Then, assigning a correct sense to a particular instance of said word (WSD)
- **Our goals**
  - Assess the effectiveness of the fusion enriched spaces
  - Evaluate the pertinence of our community discovering algorithm

Experimental Protocol

- **Feature Space**
  - Lexical (L) and Syntactic (S) Features
- **Preprocessing**
  - Remove very frequent and very infrequent words

Experimental Protocol

- **Feature Space**
  - Lexical (L) and Syntactic (S) Features
- **Preprocessing**
  - Remove very frequent and very infrequent words
- **Test Corpora**
  - Semeval Competition 2007: Train 219,554 lines. Test 50,350 lines

Experimental Protocol

- **Feature Space**
  - Lexical (L) and Syntactic (S) Features
- **Preprocessing**
  - Remove very frequent and very infrequent words
- **Test Corpora**
  - Semeval Competition 2007: Train 219,554 lines. Test 50,350 lines
- **Clustering Algorithm**
  - Spectral Clustering
  - Proposed Community Algorithm

Experimental Protocol

- **Feature Space**
  - Lexical (L) and Syntactic (S) Features
- **Preprocessing**
  - Remove very frequent and very infrequent words
- **Test Corpora**
  - Semeval Competition 2007: Train 219,554 lines. Test 50,350 lines
- **Clustering Algorithm**
  - Spectral Clustering
  - Proposed Community Algorithm
- **Evaluation Metrics**
  - Supervised Recall (SR)
  - Unsupervised F-measure (UF)

Experimental Protocol

- **Feature Space**
  - Lexical (L) and Syntactic (S) Features
- **Preprocessing**
  - Remove very frequent and very infrequent words
- **Test Corpora**
  - Semeval Competition 2007: Train 219,554 lines. Test 50,350 lines
- **Clustering Algorithm**
  - Spectral Clustering
  - Proposed Community Algorithm
- **Evaluation Metrics**
  - Supervised Recall (SR)
  - Unsupervised F-measure (UF)
  - Proposed H-measure

$$\text{H-measure} = \frac{1}{2}\left(2 * \frac{SR * UF}{SR + UF} + \frac{\delta}{\delta + |\#cl - \delta|}\right)$$

δ is the average true number of senses of the words in a test corpus

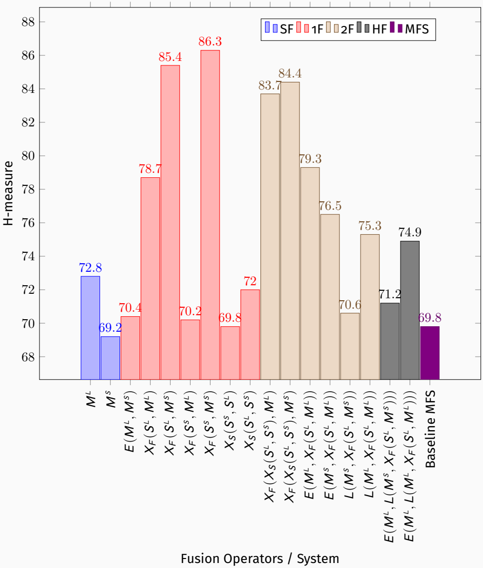
Supervised Recall

Fusion Operation / System	Recall (%) all	nouns	verbs	#cl
		Single Features		
M^L	79.20	82.10	75.80	4.13
M^S	79.10	81.60	76.20	4.47
		Early Fusion (EF)		
$E(M^L, M^S)$	78.70	81.11	76.10	4.46
		Late Cross Feature Fusion (LX_FF)		
$L(M^S, X_F(S^L, M^S))$	78.60	81.10	75.80	4.22
$L(M^L, X_F(S^L, M^L))$	79.50	82.80	75.70	3.96

Unsupervised F-measure

Fusion Operation / System	F-measure (%)			#cl
	all	nouns	verbs	
			Single Features	
M^L	72.70	76.90	67.90	4.13
M^S	69.30	69.40	69.20	4.47
			Early Fusion (EF)	
$E(M^L, M^S)$	74.00	76.66	71.11	4.46
			Cross Feature Fusion (X_FF)	
$X_F(S^S, M^L)$	78.90	80.70	76.90	1.08

Spectral Clustering Evaluation: H-measure



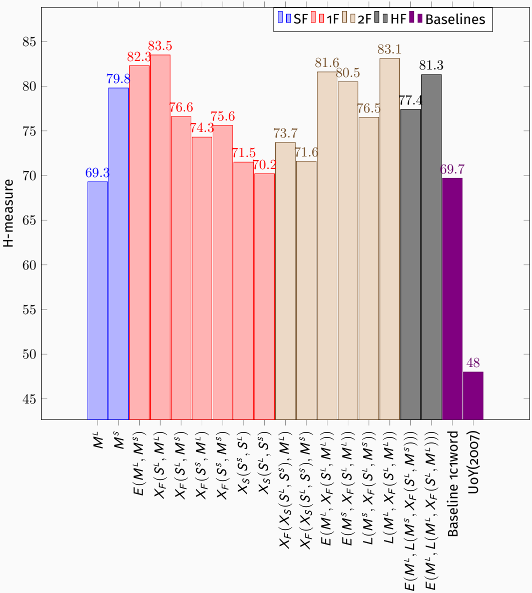
Supervised Recall

Fusion Operation / System	Recall (%)			#cl
	all	nouns	verbs	
		Single Features		
M^L	78.70	81.00	76.00	4.21
M^S	78.41	80.30	76.10	2.26
		Early Fusion (EF)		
$E(M^L, M^S)$	78.80	81.00	76.40	2.43
		Cross Feature Fusion (X_FF)		
$X_F(S^L, M^L))$	79.10	81.60	76.40	1.73

Unsupervised F-measure

Fusion Operation / System	F-measure (%)			#cl
	all	nouns	verbs	
		Single Features		
M^L	63.80	61.30.90	66.50	4.21
M^S	75.90	78.80	72.60	2.26
		Early Fusion (EF)		
$E(M^L, M^S)$	76.90	80.20	73.10	2.43
		Cross Feature Fusion (X_SF)		
$X_F(S^S, M^L)$	78.30	79.70	76.80	1.10

Proposed Algorithm Evaluation: H-measure



Conclusions

Insights From our Contributions

- **Hypergraph linguistic model to hold heterogeneous information**
    - Hypergraphs allow a multi-layered representation of text within a single resource.
    - The Wikipedia-based instantiation serves as a NLP system starting point

Insights From our Contributions

- **Hypergraph linguistic model to hold heterogeneous information**
  - Hypergraphs allow a multi-layered representation of text within a single resource.
  - The Wikipedia-based instantiation serves as a NLP system starting point
- **Multimedia fusion techniques to combine and densify representations**
  - High-degree combinations of linguistic representations reduce sparsity
  - These fusion spaces achieve improvements on NER and WSI/WSD compared to single features and trivial fusion

Insights From our Contributions

- **Hypergraph linguistic model to hold heterogeneous information**
  - Hypergraphs allow a multi-layered representation of text within a single resource.
  - The Wikipedia-based instantiation serves as a NLP system starting point
- **Multimedia fusion techniques to combine and densify representations**
  - High-degree combinations of linguistic representations reduce sparsity
  - These fusion spaces achieve improvements on NER and WSI/WSD compared to single features and trivial fusion
- **Finding semantically-related communities on linguistic networks**
  - The proposed community finding method improves over similar algorithms while being simpler and allowing for heterogeneous features

Future Work

- **Hypergraph Linguistic Model**
  - A dataframe-like structure specialized on linguistic information based on the proposed model
  - Defining inter-features similarities measures within the network

Future Work

- **Hypergraph Linguistic Model**
  - A dataframe-like structure specialized on linguistic information based on the proposed model
  - Defining inter-features similarities measures within the network
- **Combining Features and Dealing with Sparsity**
  - Finding a more principled way to determine what type of context with what type of fusion operation according to the task at hand
  - Exploring fusion with other types of features (other modalities)

Future Work

- **Hypergraph Linguistic Model**
  - A dataframe-like structure specialized on linguistic information based on the proposed model
  - Defining inter-features similarities measures within the network
- **Combining Features and Dealing with Sparsity**
  - Finding a more principled way to determine what type of context with what type of fusion operation according to the task at hand
  - Exploring fusion with other types of features (other modalities)
- **Applications to NLP**
  - Comparison with other distributional representations (word embeddings)
  - Using the large Wikipedia-based network as a background corpus to further enrich domain-specific corpora
  - Test more feature weighting schemes, validate findings on more datasets

Publications Produced by our Research

- Edmundo-Pavel Soriano-Morales, Julien Ah-Pine, Sabine Loudcher: **Fusion Techniques for Named Entity Recognition and Word Sense Induction and Disambiguation**. DS 2017
- Edmundo-Pavel Soriano-Morales, Julien Ah-Pine, Sabine Loudcher: **Using a Heterogeneous Linguistic Network for Word Sense Induction and Disambiguation.** CICLING 2016
- Edmundo-Pavel Soriano-Morales, Julien Ah-Pine, Sabine Loudcher: **Hypergraph Modelization of a Syntactically Annotated English Wikipedia Dump**. LREC 2016
- Adrien Guille, Edmundo-Pavel Soriano-Morales, Ciprian-Octavian Truica: **Topic modeling and hypergraph mining to analyze the EGC conference history**. EGC 2016
- Adrien Guille, Edmundo-Pavel Soriano-Morales: **TOM: A library for topic modeling and browsing.** EGC 2016
- Julien Ah-Pine, Edmundo-Pavel Soriano-Morales: **A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis**. DMNLP@PKDD/ECML 2016
- Sabine Loudcher, Wararat Jakawat, Edmundo-Pavel Soriano-Morales, Cécile Favre: **Combining OLAP and information networks for bibliographic data analysis: a survey**. Scientometrics 103(2) 2015

Thank you for your attention

- Edmundo-Pavel Soriano-Morales, Julien Ah-Pine, Sabine Loudcher: **Fusion Techniques for Named Entity Recognition and Word Sense Induction and Disambiguation**. DS 2017
- Edmundo-Pavel Soriano-Morales, Julien Ah-Pine, Sabine Loudcher: **Using a Heterogeneous Linguistic Network for Word Sense Induction and Disambiguation.** CICLING 2016
- Edmundo-Pavel Soriano-Morales, Julien Ah-Pine, Sabine Loudcher: **Hypergraph Modelization of a Syntactically Annotated English Wikipedia Dump**. LREC 2016
- Adrien Guille, Edmundo-Pavel Soriano-Morales, Ciprian-Octavian Truica: **Topic modeling and hypergraph mining to analyze the EGC conference history**. EGC 2016
- Adrien Guille, Edmundo-Pavel Soriano-Morales: **TOM: A library for topic modeling and browsing.** EGC 2016
- Julien Ah-Pine, Edmundo-Pavel Soriano-Morales: **A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis**. DMNLP@PKDD/ECML 2016
- Sabine Loudcher, Wararat Jakawat, Edmundo-Pavel Soriano-Morales, Cécile Favre: **Combining OLAP and information networks for bibliographic data analysis: a survey**. Scientometrics 103(2) 2015

Appendix

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:

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

    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}.

Structured Perceptron

Algorithm 1: Training phase of the Structured Perceptron

Input: Data $x \in \mathcal{X}$

Input: True labels $y \in \mathcal{Y}$

Input: Max number of iterations $MaxIteration$

Output: A vector of lerned weights w

for $Iteration = 1 \ldots MaxIterations$ **do**
  foreach $(x, y) \in \mathcal{X}, \mathcal{Y}$ **do**
    $\hat{y} = \arg\max_{\hat{y} \in \mathcal{Y}} w \cdot \Phi(x, y)$
    if $\hat{y} \neq y$ **then**
      $w \leftarrow w + \Phi(x, y) - \Phi(x, \hat{y})$
    end
  end
end
return w

The normalized Laplacian of an affinity (symmetric and positive) matrix $W \in \mathbb{R}^{n \times n}$, with $w_{ij} = w_{ji} \geq 0$, is defined as:

$$\mathcal{L} = I - D^{-1/2} W D{-1/2} \tag{1}$$

where I is the identity matrix and D is the degree matrix of W. D is defined as the diagonal matrix with the degrees $d_1, \ldots, d_n$ on the diagonal. As W may not be an adjacency matrix, we define the degrees of each row in the matrix as: $d_i = \sum_{j=1}^{n} w_{ij}$.

Given a symmetric and positive similarity matrix $W \in \mathbb{R}^{n \times n}$, and a number of desired clusters k, the steps required to perform spectral clustering are:

1. Obtain the normalized Laplacian $\mathcal{L}$ as indicated in Equation 1.
2. Obtain the first k eigenvectors $u_{1...k}$ of $\mathcal{L}$.
3. Store said eigenvectors as columns in a matrix $V \in \mathbb{R}^{n \times k}$. This matrix is akin to a lower-dimension projection of the original similarity matrix W.
4. Cluster the points in V_i with k-means. The clusters found and their members correspond to the cluster of the spectral algorithm.

Precision and Recall for NER

CONLL

Method	All_F	ALL_P	ALL_R
M^T	77.41	77.39	77.42
M^L	69.4	80.73	60.86
M^S	32.95	53.79	23.75
$E(M^T, M^L, M^S$	78.9	78.82	78.99
$E(M_L, E(E(M_T, L(M^T, X_F(S^T, M^T)))$, $L(M^L, X_F(S^S, M^L))))$	79.67	80.45	78.9

Precision and Recall for NER

WNER

Method	All_F	ALL_P	ALL_R
M^T	77.5	77.83	77.18
M^L	69.17	79.07	61.47
M^S	28.47	38.36	22.45
$E(M^T, M^L, M^S$	80.04	80.26	79.83
$E(M_L, E(E(M_T, L(M^T, X_F(S^T, M^T))),$ $L(M^L, X_F(S^S, M^L))))$	81.79	82.28	81.32

Precision and Recall for NER

WGLD			
Method	**All_F**	**ALL_P**	**ALL_R**
M^T	59.66	60.37	58.75
M^L	52.34	68.42	42.38
M^S	25.49	36.55	19.56
$E(M^T, M^L, M^S$	63.2	63.88	62.54
$E(M_L, E(E(M_T, L(M^T, X_F(S^T, M^T))), L(M^L, X_F(S^S, M^L))))$	67.05	69.63	64.64

Examples of Senses Found

Word: authority

- Dependencies
  - **process**: [neighborhood, lawyer, idea, seizure, council, subsidiary, need, collector, court, office]
  - **cabinet**: [create, trade, stability, manager, swine, department, misconduct, settlement, economist, math]
- Lexical
  - **shop**: [shop, sketch, young, month, pareo, woman, moscow, opposite, tahitian, handler, verso]
  - **supply**: [supply, justice, money, hugo, telephone, authority, initiative, alberta, bundesbank, utility, impact]
  - **evidence**: [council, machine, court, august, district, instance, fulham, auditor, hammersmith, plant]

Examples of Senses Found

Word: authority

- Best fusion operator: $X_F(S^L, M^L)$
    - **block**: [allow, including, study, told, seek, make, support, claim, provide, lawyers]
    - **veto**: [says, court, companies, years, does, law, loans, congress, trading, days]
    - **federal**: [federal, president, new, u.s., line-item, banks, local, company, airline, transportation]
    - **government**: [legislation, million, bush, year, people, billion, secretary, department, officials, house]