Yet Another Ranking Function for Automatic Multiword Term Extraction

Juan Antonio LOSSIO

Clement JONQUET

Mathieu ROCHE

Maguelonne TEISSEIRE

Laboratory of Informatics, Robotics and Microelectronics – LIRMM
fName.lName@lirmm.fr

Content

- Introduction
- Approach
- Experiments and Results
- Conclusion and Future Work

Introduction

SIFR Project

Funded

International Collaboration

- Pr. Mark A. Musen, MD, PhD (BMIR, Stanford University).

Introduction: Context

- Creating and maintaining terminologies

Introduction: Context

- Creating and maintaining terminologies

expensive task if done by human experts.

Introduction: Context

- Creating and maintaining terminologies

expensive task if done by human experts.

. Life Sciences,
. Medicine and
. Biology

Most active areas for the development of terminology resources.

Introduction: Context

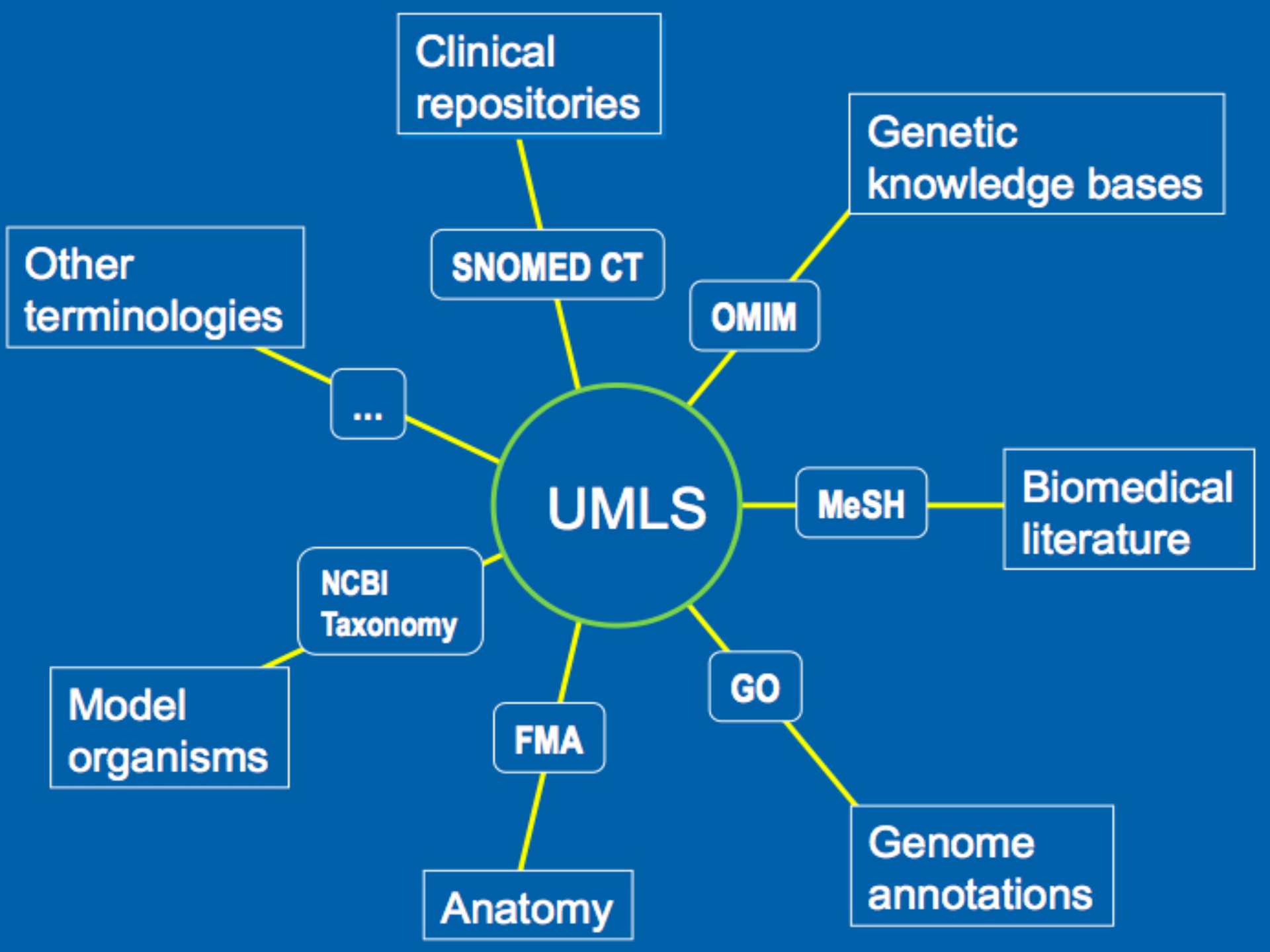
- Creating and maintaining terminologies

expensive task if done by human experts.

. Life Sciences,
. Medicine and
. Biology

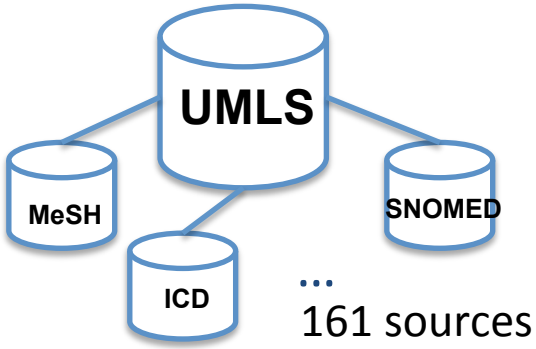
Most active areas for the development of terminology resources.

Unified Medical Language System


Clinical repositories
Genetic knowledge bases
Other terminologies
SNOMED CT
OMIM
...
UMLS
MeSH
Biomedical literature
NCBI Taxonomy
Model organisms
FMA
GO
Anatomy
Genome annotations

Introduction: Context

Unified Medical Language System



~ 6 000 000 terms for English

Introduction: Context

Table 1: Overview of terminological resources in French with links to the UMLS (2012AA) that are publicly available

Source vocabulary	Number of Strings	Number of CUIs
UMLS 2012AA	171,764	85,685
MSHFRE	105,758	41,229
MDRFRE	65,071	48,005
WHOFRE	3,631	3,091
MTHMSTFRE	1,833	1,636
ICPCFRE	702	722
ICD10[1]	26,337	12,143
FMA[2]	4,564	4,452
SNOMED[3]	139,792	93,632
ICNP[4]	2,801	1,158
All	336,264	169,123

Source: ***Language Resources for French in the Biomedical Domain.*** Aurelie Neveol, Julien Grosjean, Stéfan Darmoni and Pierre Zweigenbaum. **LREC 2014**

Introduction: Context

Table 1: Overview of terminological resources in French with links to the UMLS (2012AA) that are publicly available

Source vocabulary	Number of Strings	Number of CUIs
UMLS 2012AA	171,764	85,685
MSHFRE	105,758	41,229
MDRFRE	65,071	48,005
WHOFRE	3,631	3,091
MTHMSTFRE	1,833	1,636
ICPCFRE	702	722
ICD10[1]	26,337	12,143
FMA[2]	4,564	4,452
SNOMED[3]	139,792	93,632
ICNP[4]	2,801	1,158
All	336,264	169,123

Source: ***Language Resources for French in the Biomedical Domain.*** Aurelie Neveol, Julien Grosjean, Stéfan Darmoni and Pierre Zweigenbaum. **LREC 2014**

Introduction: Context

Table 3: Overview of French-only terminological resources

Source vocabulary	Number of (French) Strings	Number of Native Concepts	Number of CUIs mapped	Availability
BNPC[1]	91,750	103,280	30,564	proprietary
CCAM	9,666	18,314	7	browsing in HeTOP
ADICAP	9,189	9,189	143	browsing in HeTOP
Cladimed[2]	4,546	4,548	191	proprietary
LPP[3]	4,546	4,546	0	browsing in HeTOP
DRC[4]	3,313	3,324	520	browsing in HeTOP
BHN[5]	2,534	2,544	0	browsing in HeTOP
NABM	1,084	1,084	0	browsing in HeTOP
BNCI[6]	786	802	351	proprietary

Source: ***Language Resources for French in the Biomedical Domain.*** Aurelie Neveol, Julien Grosjean, Stéfan Darmoni and Pierre Zweigenbaum. **LREC 2014**

Introduction: Context

Résumé :

- Expensive task done by human experts
- Nb of terms FR < Nb of terms EN

  ~300 000 < ~6 000 000

- No availabitly of French resources
- Hard collaboration with other teams

Introduction

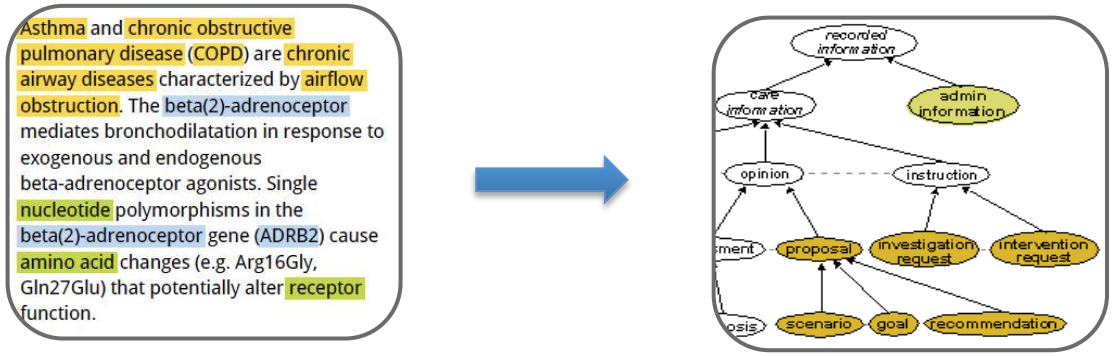
Objectives

Construction and enrichment of French biomedical terminology

How

- Extracting and linking of all information
- Enriching/constructing ontologies

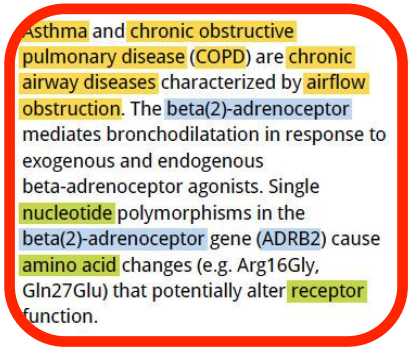
Introduction

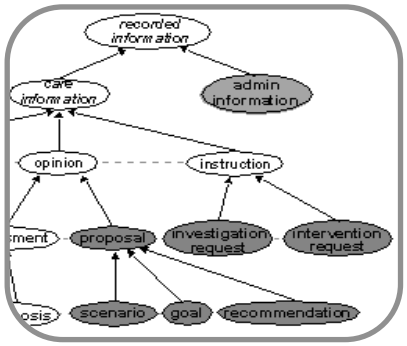


Biomedical Term Extraction

Ontology Enrichment

Introduction





Biomedical Term Extraction

Ontology Enrichment

Introduction

Extraction of biomedical terms from free text

Motivation

Few studies related to French Automatic Term Extraction

Challenge

- Processing of French biomedical data
- Transforming heterogeneous data in homogeneous data

Content

- Introduction
- Approach
- Experiments and Results
- Conclusion and Future Work

Approach


Linguistic and Statistic
Graph
Text
Term ranking 1
LIDF-value
Term Re-ranking 2
TeRGraph
term1
term2
...
termn

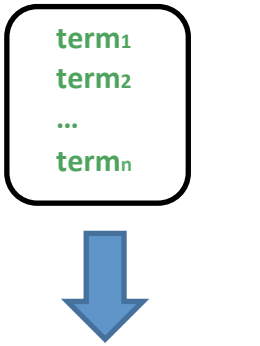
Approach


Linguistic and Statistic
Graph
Text
Term ranking 1
LIDF-value
Term Re-ranking 2
TeRGraph
term1
term2
...
termn

Approach

A) Linguistic measure:

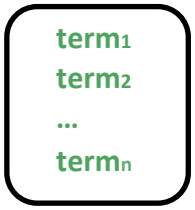
Input: List of UMLS terms:



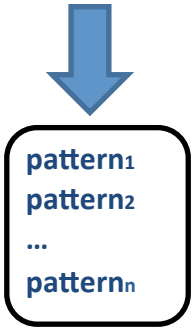
Approach

A) Linguistic measure:

Input: List of UMLS terms:



Build Biomedical Patterns



Output: List of patterns.

NN is a noun, NP a proper noun, JJ an adjective, CD a cardinal number, POS a possessive ending

Pattern	Frequency	Probability
NP POS NN	*3200*	*3200/7090 =* ***0.45***
NN NN NP CD	*1430*	*1430/7090 =* ***0.20***
JJ JJ NN CD	*810*	*810/7090 =* ***0.12***
JJ NN NP POS NN	*1650*	*1650/7090 =* ***0.23***
	7090	***1.00***

Approach

B) Statistic measure:

1. Part-of-Speech Tagging
2. Candidate term extraction
3. Ranking of candidate terms

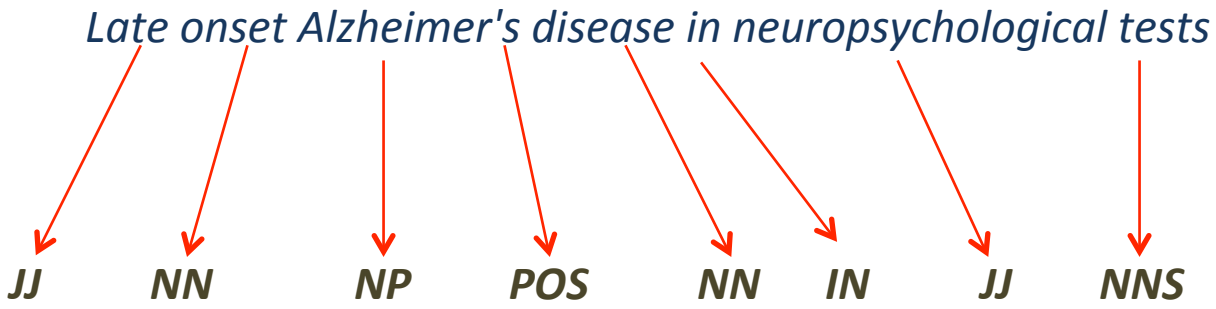
Approach

B) Statistic measure:

1. **Part-of-Speech Tagging**
2. Candidate term extraction
3. Ranking of candidate terms

(1) Part-of-Speech Tagging

Assign each word in a text to its grammatical category (e.g., noun, adjective).



We apply part-of-speech to the whole corpus

Three tools:

- **Stanford Tagger,**
- TreeTagger,
- Brill's rules

Approach

B) Statistic measure:

1. Part-of-Speech Tagging
2. **Candidate term extraction**
3. Ranking of candidate terms

(2) Candidate Term Extraction

Early onset Alzheimer's disease in neuropsychological tests

JJ ***NN*** ***NP*** ***POS*** ***NN*** ***IN*** ***JJ*** ***NNS***

Pattern	Frequency	Probability
NP POS NN	*3200*	*3200/7090 =* ***0.45***
NN NN NP CD	*1430*	*1430/7090 =* ***0.20***
JJ JJ NN CD	*810*	*810/7090 =* ***0.12***
JJ NN NP POS NN	*1650*	*1650/7090 =* ***0.23***
	7090	***1.00***

(2) Candidate Term Extraction

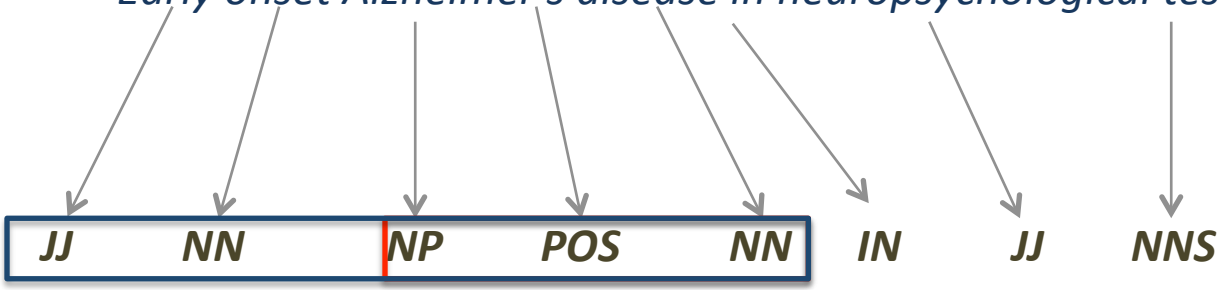
Early onset Alzheimer's disease in neuropsychological tests

JJ ***NN*** ***NP*** ***POS*** ***NN*** ***IN*** ***JJ*** ***NNS***

Pattern	Frequency	Probability
NP POS NN	*3200*	*3200/7090 =* ***0.45***
NN NN NP CD	*1430*	*1430/7090 =* ***0.20***
JJ JJ NN CD	*810*	*810/7090 =* ***0.12***
JJ NN NP POS NN	*1650*	*1650/7090 =* ***0.23***
	7090	***1.00***

(2) Candidate Term Extraction

Early onset Alzheimer's disease in neuropsychological tests



Pattern	Frequency	Probability
NP POS NN	*3200*	*3200/7090 =* ***0.45***
NN NN NP CD	*1430*	*1430/7090 =* ***0.20***
JJ JJ NN CD	*810*	*810/7090 =* ***0.12***
JJ NN NP POS NN	*1650*	*1650/7090 =* ***0.23***
	7090	***1.00***

(2) Candidate Term Extraction

Early onset Alzheimer's disease in neuropsychological tests

Pattern	Frequency	Probability
NP POS NN	*3200*	*3200/7090 =* ***0.45***
NN NN NP CD	*1430*	*1430/7090 =* ***0.20***
JJ JJ NN CD	*810*	*810/7090 =* ***0.12***
JJ NN NP POS NN	*1650*	*1650/7090 =* ***0.23***
	7090	***1.00***

Alzheimer's disease

Early onset Alzheimer's disease

Approach

B) Statistic measure:

1. Part-of-Speech Tagging
2. Candidate term extraction
3. **Ranking of candidate terms**

Approach

B) Statistic measure:

$$LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value$$

Approach

C-value: **Statistical Methods**

C-value: *Improves the extraction of longest terms*

soft contact **soft contact lens**

Source : ***Automatic recognition of multi-word terms: the C- value/NC-value Method***. Frantzi K., Ananiadou S., Mima, H.:. International Journal on Digital Libraries, vol. 3, pp. 115-130, (2000)

Approach

C-value: **Statistical Methods**

C-value: *Improves the extraction of longest terms*

soft contact | **soft contact lens**

Source : ***Automatic recognition of multi-word terms: the C- value/NC-value Method***. Frantzi K., Ananiadou S., Mima, H.:. International Journal on Digital Libraries, vol. 3, pp. 115-130, (2000)

Approach

B) Statistic measure:

$$LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value$$

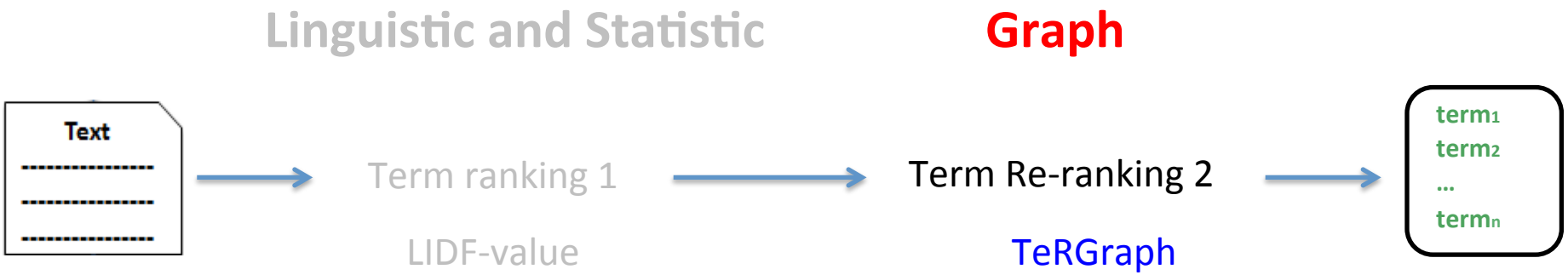
C-value

$$w(a) \times \left(f(a) - \frac{1}{|S_a|} \times \sum_{b \in S_a} f(b) \right)$$

1) *Alzheimer's disease*
2) *Early onset Alzheimer's disease*

$$w(a) \times f(a) \quad \text{if } a \notin nested$$

Approach


Linguistic and Statistic
Graph
Text
Term ranking 1
LIDF-value
Term Re-ranking 2
TeRGraph
term1
term2
...
termn

Approach

TeRGraph: *Terminology Ranking based on Graph information*

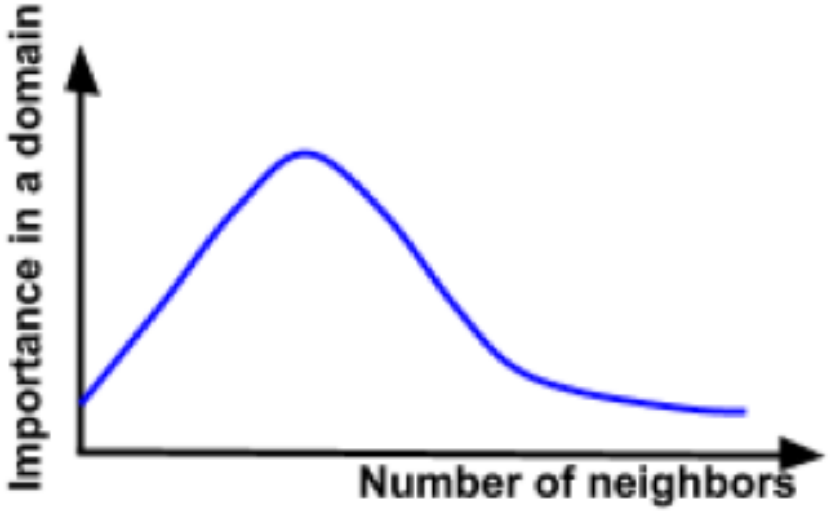
Vertices = multiword terms
Edges = co-occurrence between multiword terms in the sentences in the corpus.

Approach

TeRGraph: *Terminology Ranking based on Graph information*

Vertices = multiword terms
Edges = co-occurrence between multiword terms in the sentences in the corpus.

Hypotheses: A term with more neighbors is less representative for a domain. This means that this term is used in the general domain.



Approach

TeRGraph: *Terminology Ranking based on Graph information*

Vertices = multiword terms
Edges = co-occurrence between multiword terms in the sentences in the corpus.

Approach

TeRGraph: *Terminology Ranking based on Graph information*

Vertices = multiword terms
Edges = co-occurrence between multiword terms in the sentences in the corpus.

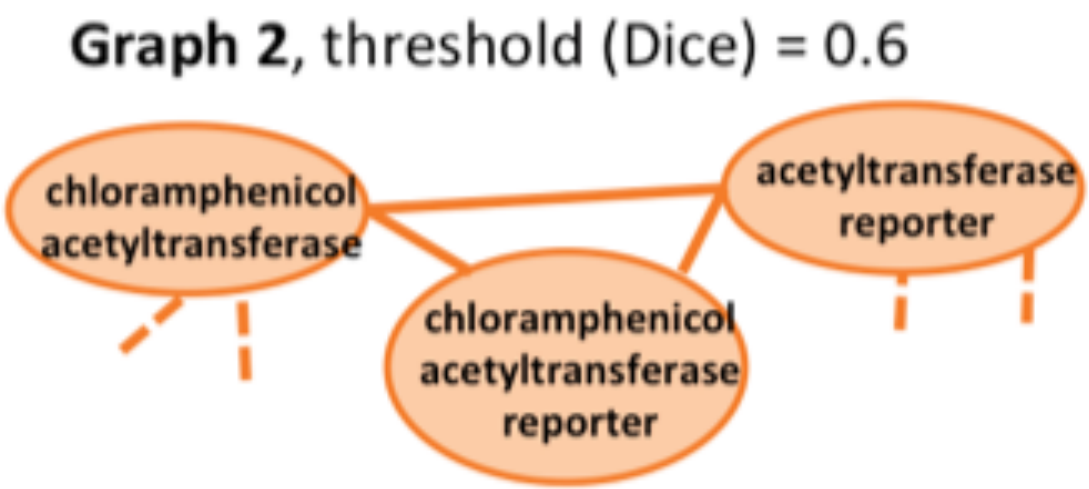
$$D(x,y) = \frac{2 \times P(x,y)}{P(x) + P(y)}$$

Approach

TeRGraph: *Terminology Ranking based on Graph information*

Vertices = multiword terms
Edges = co-occurrence between multiword terms in the sentences in the corpus.

$$D(x,y) = \frac{2 \times P(x,y)}{P(x) + P(y)}$$



Approach

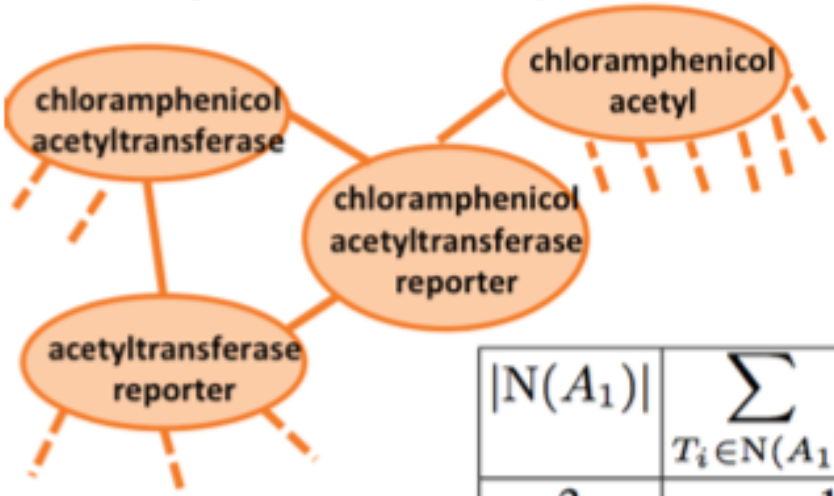
Representativeness computations on the term graph: re-rank the list of extracted terms

$$TeRGraph(A) = \log_2 \left(1.5 + \frac{1}{|\mathrm{N}(A)| + \sum_{T_i \in \mathrm{N}(A)} |\mathrm{N}(T_i)|} \right)$$

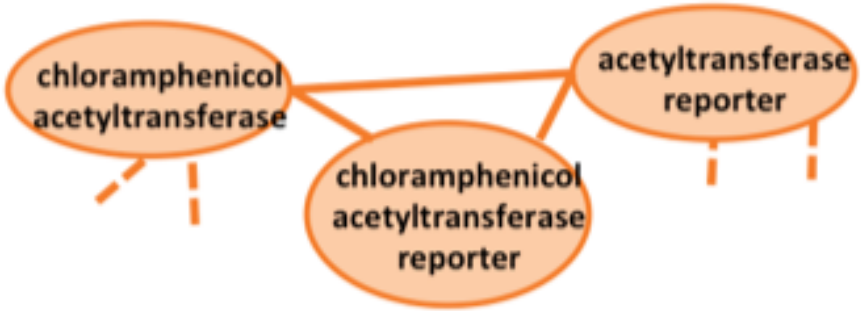
Approach

Representativeness computations on the term graph: re-rank the list of extracted terms

Graph 1, threshold (Dice) = 0.5



Graph 2, threshold (Dice) = 0.6



$\lvert N(A_1)\rvert$	$\sum_{T_i \in N(A_1)} \lvert N(T_i)\rvert$	$\lvert N(A_2)\rvert$	$\sum_{T_j \in N(A_2)} \lvert N(T_j)\rvert$
3	16	2	8
$TeRGraph(A_1) = 0.632$		$TeRGraph(A_2) = 0.670$	

Approach


Linguistic and Statistic
Graph
Text
Term ranking 1
LIDF-value
Term Re-ranking 2
TeRGraph
term1
term2
...
termn

Content

- Introduction
- Approach
- Experiments and Results
- Conclusion and Future Work

Experiments and Results

Dataset and Protocol

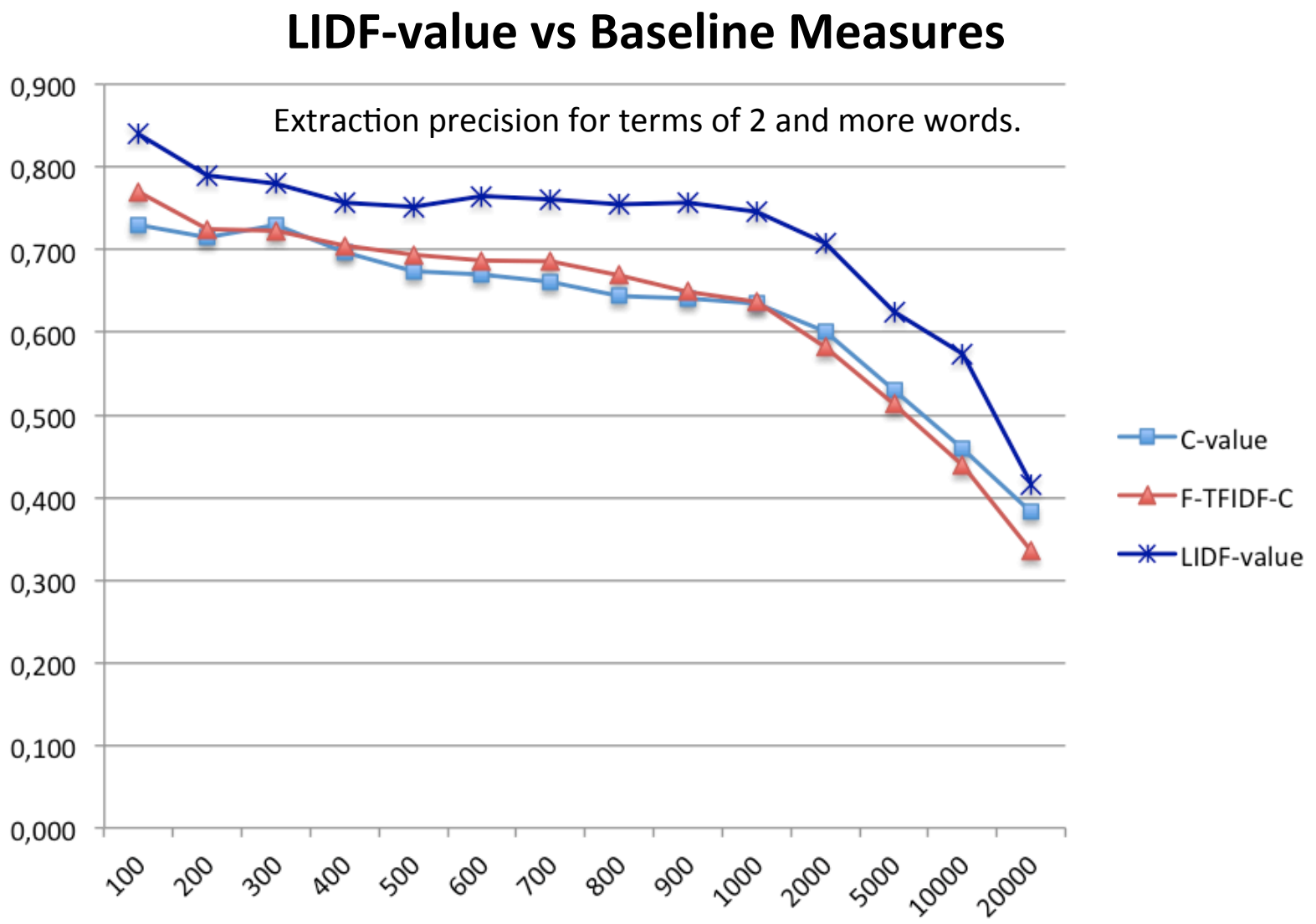
GENIA		(EN) = 400 000 words.

Composed of 2 000 titles and abstracts of journal articles that have been taken from the **Medline**

- molecular biology,
- proteins,
- genes,
- cells

Precision: P@k

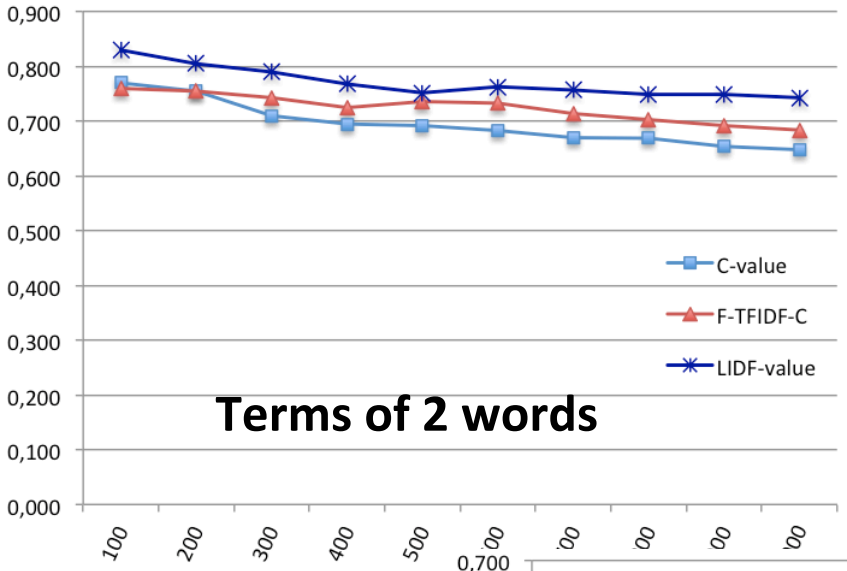
Results:

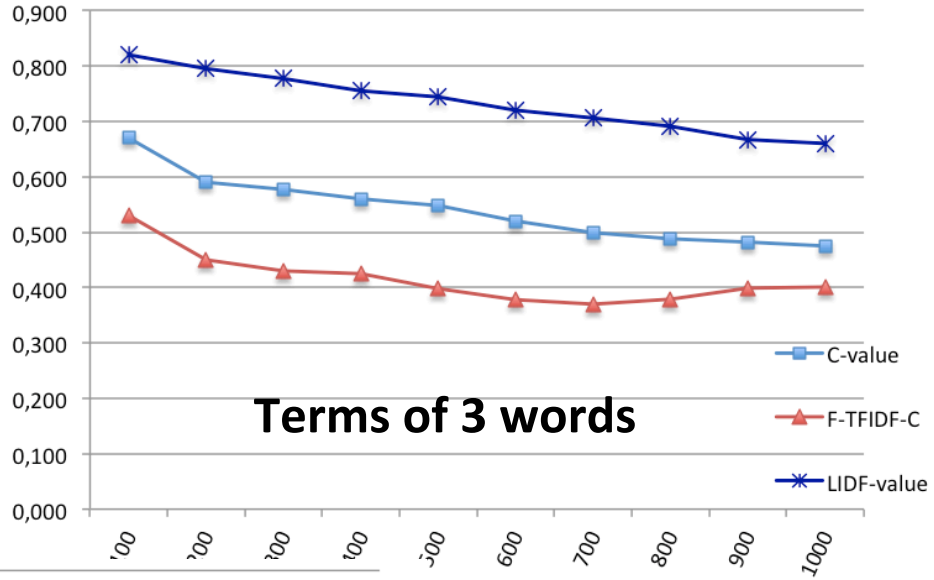

LIDF-value vs Baseline Measures
Extraction precision for terms of 2 and more words.
0,900
0,800
0,700
0,600
0,500
0,400
0,300
0,200
0,100
0,000
100
200
300
400
500
600
700
800
900
1000
2000
5000
10000
20000
C-value
F-TFIDF-C
LIDF-value

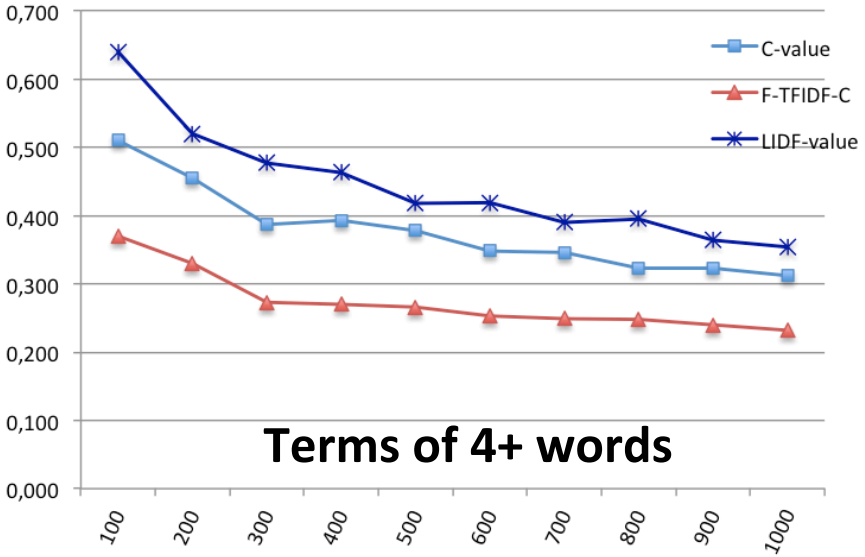
Results

Extraction precision:

LIDF-value vs Baseline Measures







Results:

TeRGraph

	$\delta \geq 0.25$	$\delta \geq 0.35$	$\delta \geq 0.50$	$\delta \geq 0.60$	$\delta \geq 0.70$
P@100	0.840	0.860	0.910	**0.930**	0.900
P@200	0.800	0.790	0.850	**0.855**	**0.855**
P@300	0.803	0.773	0.833	**0.830**	0.820
P@400	0.780	0.732	**0.820**	**0.820**	0.815
P@500	0.774	0.712	0.798	**0.810**	0.806
P@600	0.773	0.675	0.797	**0.807**	0.792
P@700	0.760	0.647	0.769	**0.796**	0.787
P@800	0.756	0.619	0.748	**0.784**	0.779
P@900	0.748	0.584	0.724	0.773	**0.777**
P@1000	0.751	0.578	0.720	0.766	**0.769**
P@2000	0.689	0.476	0.601	0.657	**0.694**
P@3000	0.642	0.522	0.535	0.605	**0.644**
P@4000	**0.612**	0.540	0.543	0.559	0.593
P@5000	**0.574**	0.546	0.544	0.554	0.562
P@6000	0.558	0.539	0.540	0.549	**0.561**
P@7000	**0.556**	0.540	0.540	0.545	0.552
P@8000	**0.546**	**0.546**	**0.546**	**0.546**	**0.546**

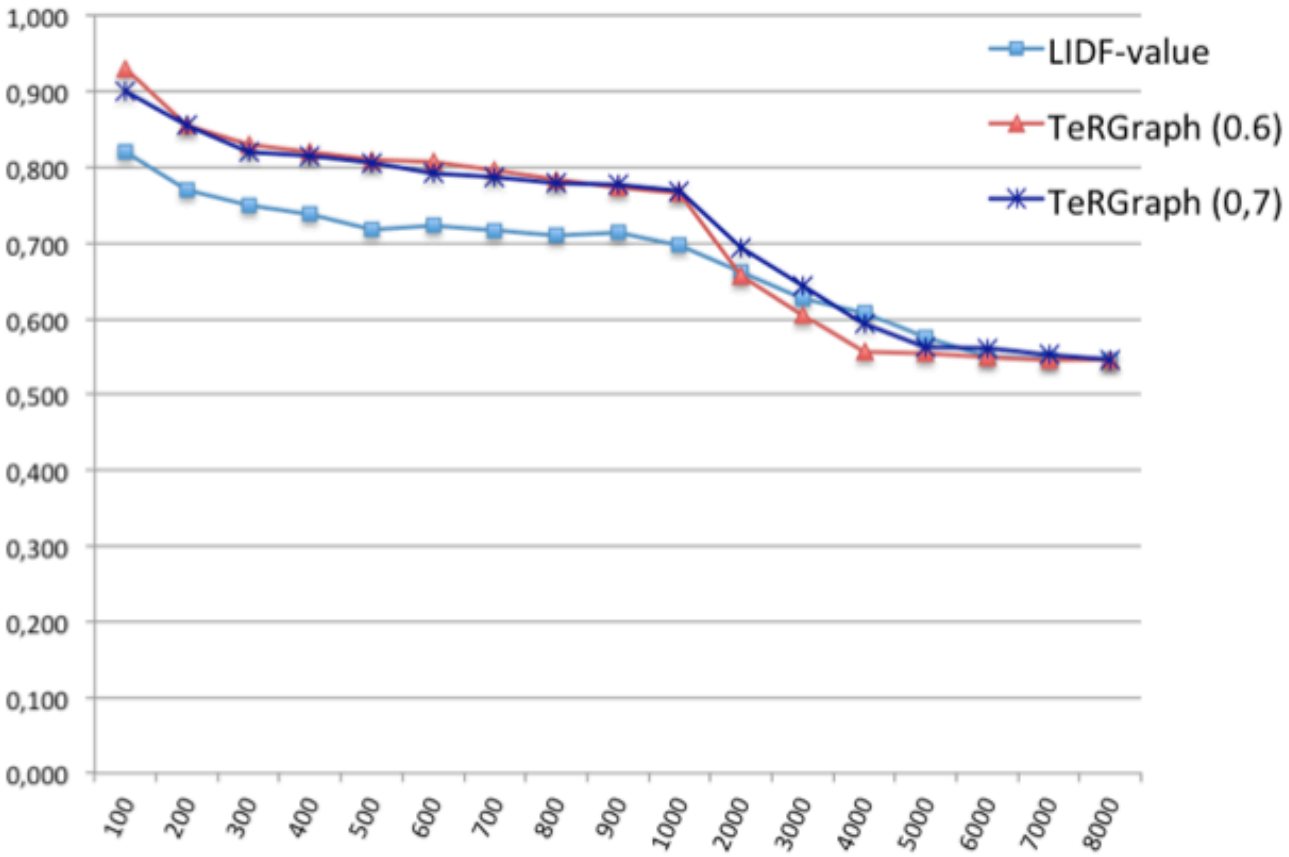
Results:

TeRGraph

	$\delta \geq 0.25$	$\delta \geq 0.35$	$\delta \geq 0.50$	$\delta \geq 0.60$	$\delta \geq 0.70$
P@100	0.840	0.860	0.910	**0.930**	0.900
P@200	0.800	0.790	0.850	**0.855**	**0.855**
P@300	0.803	0.773	0.833	**0.830**	0.820
P@400	0.780	0.732	**0.820**	**0.820**	0.815
P@500	0.774	0.712	0.798	**0.810**	0.806
P@600	0.773	0.675	0.797	**0.807**	0.792
P@700	0.760	0.647	0.769	**0.796**	0.787
P@800	0.756	0.619	0.748	**0.784**	0.779
P@900	0.748	0.584	0.724	0.773	**0.777**
P@1000	0.751	0.578	0.720	0.766	**0.769**
P@2000	0.689	0.476	0.601	0.657	**0.694**
P@3000	0.642	0.522	0.535	0.605	**0.644**
P@4000	**0.612**	0.540	0.543	0.559	0.593
P@5000	**0.574**	0.546	0.544	0.554	0.562
P@6000	0.558	0.539	0.540	0.549	**0.561**
P@7000	**0.556**	0.540	0.540	0.545	0.552
P@8000	**0.546**	**0.546**	**0.546**	**0.546**	**0.546**

Results:

Precision comparison of *LIDF-value* and *TeRGraph*



Content

- Introduction
- Approach
- Experiments and Results
- Conclusion and Future Work

Conclusion and Future Work

- Two new measures for multi-word term extraction
- **LIDF-value**: Linguistic and Statistic measure
- **TeRGraph**: Graph-based measure
- These outperform the state-of-the-art reference measures

Future Work

- Enrich our dictionaries
- Improve results with web-based measure
- Test this approach on other domains: ecology and agronomy
- Experiment our proposals on French and Spanish corpora