Semantic metadata annotation.
Tagging Medline abstracts for enhanced information access

Fidelia Ibekwe-SanJuan
Elico
Université de Lyon 3

Background

- Corporate organizations and people need more fine-grained access to electronic content
- Semantic retrieval solutions are flourishing (*iSeek*, *SmartLogic*, *Cogito*)
- Semantic components are embedded into corporate versions of popular search tools (*Google*, *MicroSoft*)

Background

Semantic metadata can be:

- broad thematic categories (sports, news, politics,travel, ...)
- named entities (persons, places, brand names, ...)
- structural roles (actors, agents, instruments, ...)
- argumentative roles (aims, method, results, conclusions, ...)
- Domain-specific entities (genes, proteins,...)
- etc...

Previous studies

- Science → problem solving activity (Swales 1990, Tbahriti et al. 2005)
- Scientific articles follow a relatively fixed argumentative structure (Swales 1990; Salager-Meyer 1992; Halliday & Martin 1993; Teufel & Moens 1999 & 2002)
- This structure is materialized by the presence of certain divisions/sections
- The 4 major ones: aim – method – result - conclusion

Previous studies

The experimental sciences follow these divisions more
Professionnel guide ANSI/NISO Z39.14-1979

- Abstracts → same argumentative divisions as full texts
- The linguistic cues that announce them are neither systematic nor fixed in their form
- introduction and conclusion sentences → important for judging a paper's worth *wrt* to a topic (Teufel & Moens 2002, Ruch *et al*. 2007)

Objectives

- Categorize sentences by type of argumentative information
- These are high level categories of information that are largely common to scientific discourse
- Not really dependent on a particular scientific domain

Possible applications

- Automatique summarization (Luhn 1958, DeJong 1982, Teufel & Moens 1999, Orasan 2001)
- Semantic information retrieval (Ruch *et al.* 2007, Ibekwe-SanJuan *et al.* 2008)

"*get me records where **method X** is used to cure symptom Y*"

- Information extraction (biomedical domain)
- Novelty detection (Mizuta et al. 2005, McNight & Srinivasan 2003, Lisacek et al. 2005)

Corpus

- Experimental Sciences : 50 abstracts from pre-prints in Quantitative Biology (Ibekwe-SanJuan 2005)
- Humanities : 1000 abstracts in Information Science from PASCAL/ INIST database (Ibekwe-SanJuan 2005)
- Physics: 1293 abstracts in Astronomy (Ibekwe-SanJuan et al. 2008)
- 200 abstracts from Medline with argumentative divisions

Sentence Categorization

Choice of abstracts (concision, no or few anaphora, lexical compactness, key sentences, key contributions)

1. Objective
2. Result
3. Related work
4. Newthing (discovery)
5. Future work
6. Hypothesis
7. Conclusion
8. Method (not modeled linguistically)

Methodology

Test two approaches to sentence categorization :

- Linguistic cues : lexico-syntactic patterns
- Pseudo-statistique : positional cues

Linguistic cues

- cue words and lexico-syntactic patterns
- Non content bearing words *(not those used for indexing*)
- Necessitates lexical, morphological and a little syntax information

Sample linguistic cues

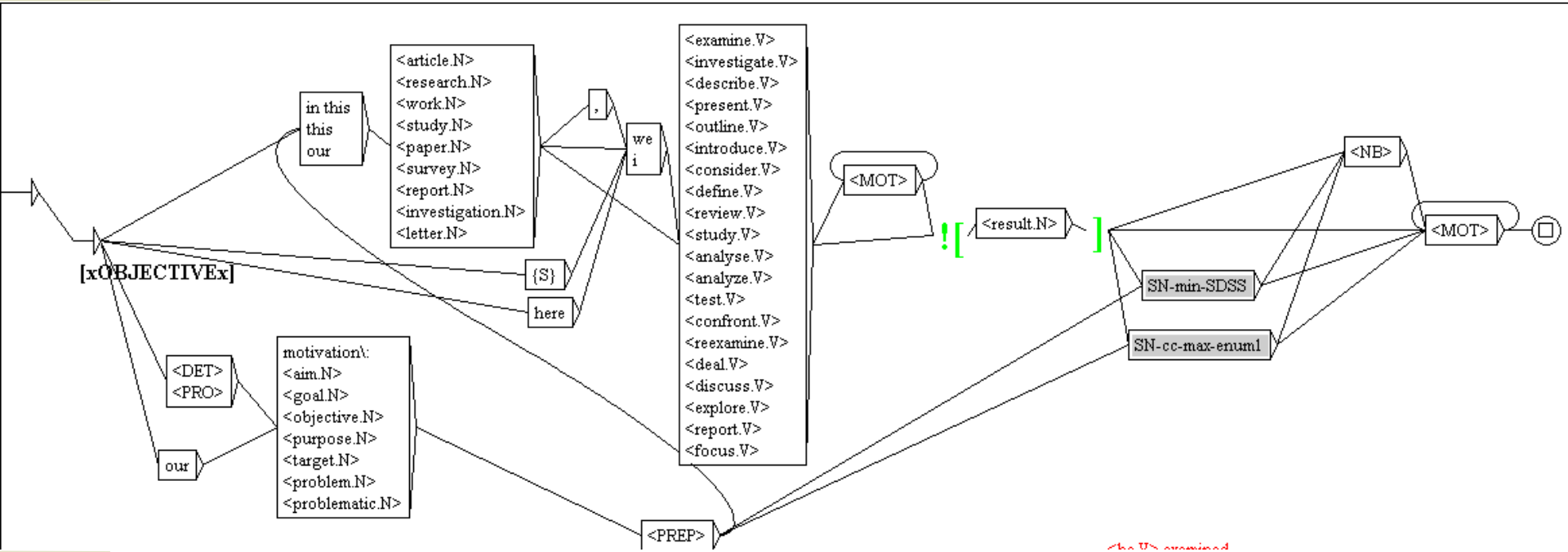
OBJECTIVE	In this_{current|present} {article|paper|study|research|work}... We_{examine|investigate|describe|present|outline|introduce}.... DET_{motivation:|aim|goal|objective|problem}...
NEWTHING	we propose a novel approach | This analysis reveals | Emerging evidence suggests that | Interestingly, our results indicate that
RELATED_WORK	{in contrast to|unlike|in common|n comparison to|in contrast to common belief|despite} our {work|study|hypothesis|observation|approach..} {<contradict.V>|<disagree.V>...}
RESULT	In this paper we show that | Our research suggests that | Results confirm that | It is shown here for the first time that
HYPOTHESIS	DET_NP_{may|might}_{ADV|V_NP} | Our findings support the view that
FUTURE_WORK	{Further|Future|more}_{work|investigation|observation}_<verb>
CONCLUSION	This paper concludes ... |Conclusion: |Finally, ... |As a conclusion

Linguistic cues

- Implemented as finite state automata in the Unitex tool (S. Paumier, Université de Marne la Vallée)
- Certain rules are embedded in others
- Applied on Medline abstracts stripped of their original title of argumentative divisions

Linguistic cues

Grammar for categorizing « objective » sentences



Linguistic cues

Concordances for « objective » sentences

[xOB]ECTIVEx] Our objective was to
[xOB]ECTIVEx] Our objectives were to
[xOB]ECTIVEx] our study was to establish whether Helicobacter pylori treatment has any influence on the level of soluble form of MUC 1 mucin in gastric juice
[xOB]ECTIVEx] our study was to investigate immunohistochemical, stereological and ultrastructural changes of rat K cells after
[xOB]ECTIVEx] The goal of screening is to detect clinically significant prostate cancers at a stage when intervention reduces morbidity and mortality
[xOB]ECTIVEx] The goal of this study was to investigate the expression pattern of CXCL13 in comparison with PD
[xOB]ECTIVEx] the goal of total eradication of all visible tumor formations
[xOB]ECTIVEx] the objective of
[xOB]ECTIVEx] The objective of this investigation was to prove that Al from an organic metal complex is able to activate GS
[xOB]ECTIVEx] The objective of this study was to demonstrate that fish-processing by-products could be used as sole raw material to sustain the growth of Staphylococcu
[xOB]ECTIVEx] The objective of this study was to determine whether the F508C variant in the cystic fibrosis transmembrane conductance regulator gene has a significant
[xOB]ECTIVEx] The objective of this study was to investigate the effects and the regulatory mechanism of CGRP on the expression and activity of nitric oxide synthase (N
[xOB]ECTIVEx] The objective of this study was to isolate
[xOB]ECTIVEx] The objective of this work was to analyze whether hsp-20 and rap-1a are expressed in sexual stages and kinetes of Babesia bigemina
[xOB]ECTIVEx] The objective of this work was to clone
[xOB]ECTIVEx] The objective of this work was to study the immunogenic capacity of a 156-kDa recombinant protein of Clostridium chauvoei that has shown
[xOB]ECTIVEx] The present study explores the action of midazolam on endothelial activation and its role to peripheral benzodiazepine receptor (PBR) in cultured human
[xOB]ECTIVEx] The present study was conducted to assess the influence of these factors on transplantation-associated injury independently or in combination
[xOB]ECTIVEx] The present study was designed to investigate whether use of left ventricular assisted technique (LVA) in beating
[xOB]ECTIVEx] the present study was to assess adrenocortical function in RA females
[xOB]ECTIVEx] the present study was to investigate the activity of tyrosine hydroxylase

Positional cues

- Based on empirical observations:
  - argumentative divisions follow one another in an orderly & logical fashion
  - this criterion seemed to work rather well on Medline abstracts

Positional cues (Heuristics)

From beginning of abstract:

- up to ¼ of sentences = objective
- after ¼ and before ½ of sentences = method
- between ½ - ¾ of sentences = result
- from ¾ until the last sentence = conclusion

Positional cues: have a 100% coverage of sentences in an abstract

Evaluation

- Compare accuracy of annotation by linguistic et positional cues
- Against original argumentative divisions in MEDLINE
- However, great variability in division names
  - 35 different division names in 200 abstracts
- MEDLINE does not annotate each sentence
  - ach division preceded by its name
  - before evaluation, we need to propagate the division name to all its sentences

Argumentative division names in MEDLINE

Division names found in 200 MEDLINE abstracts	Mapped to
OBJECTIVE, BACKGROUND, INTRODUCTION, AIM, AIMS, AIMS AND BACKGROUND, BACKGROUND/AIMS, CONTEXT, PURPOSE	**OBJECTIVE**
METHOD, METHODS, METHODOLOGY, DESIGN, DESIGN, SETTING AND PARTICIPANTS, STUDY DESIGN, STUDY DESIGN AND METHODS, RESEARCH DESIGN, PATIENTS AND METHODS, MATERIAL AND METHOD, MATERIALS AND METHODS, DATA SOURCE, DATA SUMMARY, SETTING, PARTICIPANTS, SUBJECTS, PARTICIPANTS, INTERVENTION	**METHOD**
RESULTS, FINDINGS, METHODS AND RESULTS, METHODOLOGY/PRINCIPAL FINDINGS, OUTCOME MEASURES, MAIN OUTCOME MEASURE, MAIN OUTCOME MEASURES	**RESULT**
LIMITATIONS, INTERPRETATION, CONCLUSION, CONCLUSIONS, DISCUSSION, CONCLUSIONS/SIGNIFICANCE, SIGNIFICANCE AND IMPACT OF THE STUDY	**CONCLUSION**

Evaluation

- Ruch et al. (2007) used ML techniques to learn the 4 major argumentative roles in MEDLINE sentences:

  objective – méthode – résultat – conclusion

- To be comparable, our study needs to map our 7 linguistic annotations into the same 4 classes:

  objective → objective

  result, related_work, newthing → result

  hypothesis, future_work, conclusion→ conclusion

  Méthode → not modeled by linguistic cues

Evaluation

- *Precision* : proportion of correctly annotated sentences of a certain type
- *Recall* : proportion of annotated sentences over all sentences of type *x*.

- F-score

$$F = \frac{2.\ \text{precision}\ .\ \text{recall}}{(\text{precision} + \text{recall})}$$

Evaluation

		Linguistic cues				
Medline tags	nb	Agree	total	Prec	Recall	F-score
Obj	417	34	53	0.64	0.08	0.14
Method	439	-	-	-	-	-
Results	713	8	218	0.04	0.01	0.02
Conclusion	301	92	157	0.59	0.31	0.40
Total	*1870*	*134*	*428*	*0.31*	*0.07*	*0.12*

Total sentences

Evaluation

				Positional cues		
Medline tags	nb	Agree	total	Prec.	Recall	F-score
Obj	417	333	409	0.81	0.80	0.81
Method	439	244	408	0.60	0.56	0.58
Results	713	336	486	0.69	0.47	0.56
Conclusion	301	288	567	0.51	0.96	0.66
Total	*1870*	*1201*	*1870*	*0.64*	*0.64*	*0.64*

Findings...

- Poor performance of linguistic cues on Medline abstracts
- The cues were often absent, hence poor sentence coverage
- Linguistic cues are subject to variations in their form
- OBJECTIVE and CONCLUSION : easier to categorize (Ruch et al. 2007, Teufel & Moens 2002)
- Difficulty in detecting METHOD sentences
- Positional cues are useful in boosting performance by another method (ML, Ruch et al. 2007)

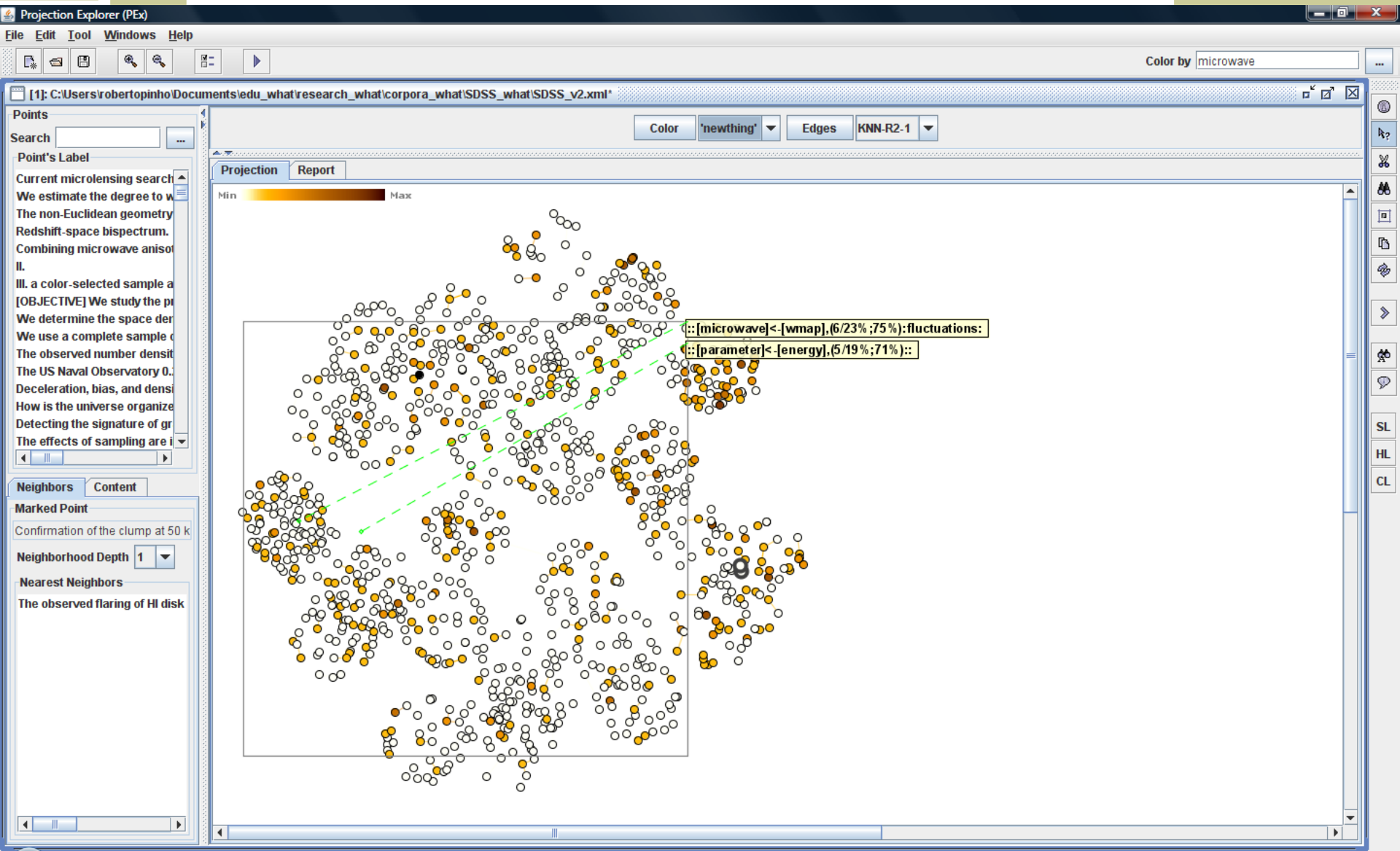
Perspectives

- Semantic IR

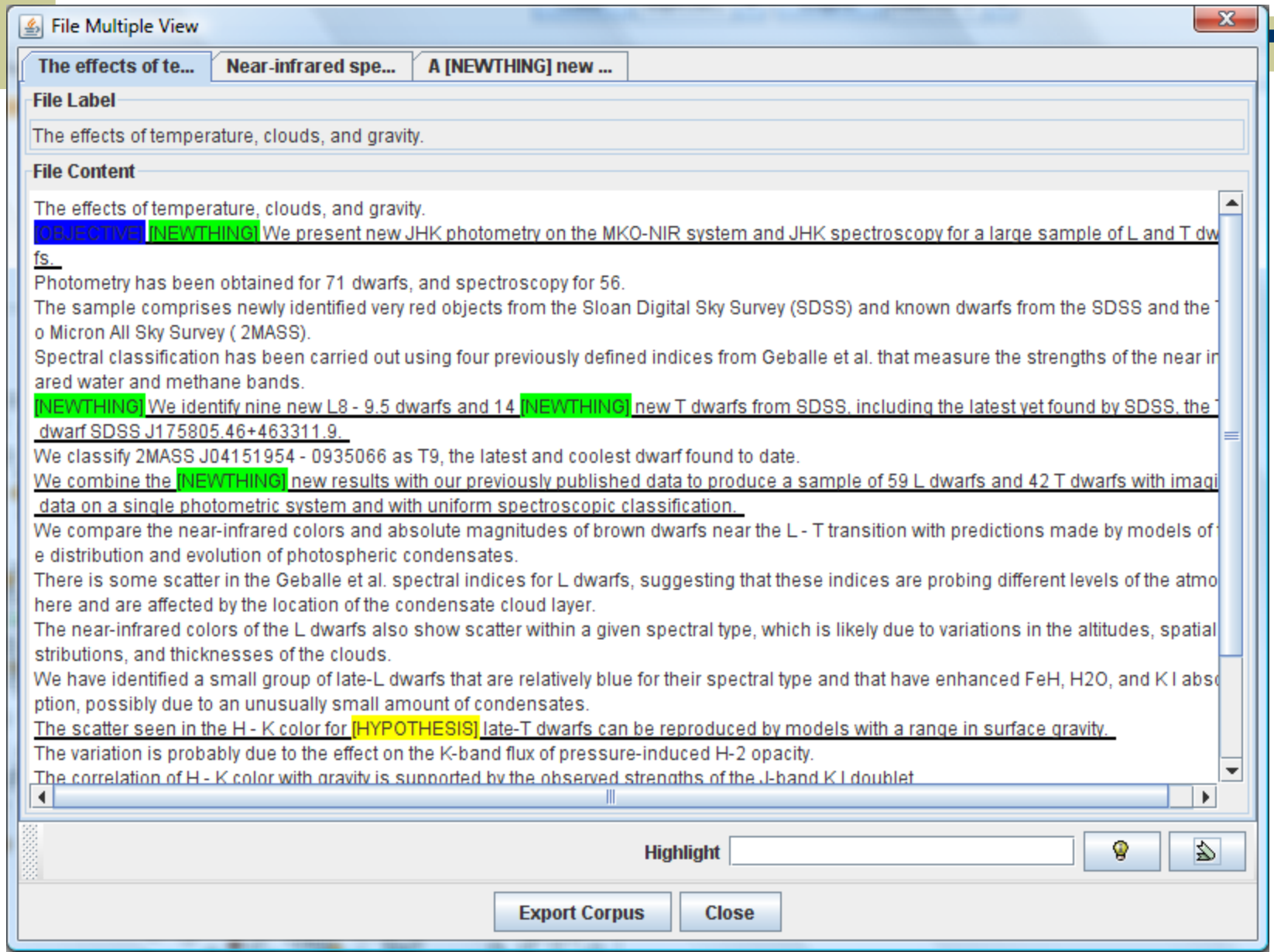
Query-oriented multi-abstract summarization using semantic annotations (Ibekwe-SanJuan et al. 2008)

- Information visualisation (Ibekwe-SanJuan et al. 2008)
- Novelty detection

Perspectives



Perspectives



Thanks!

….Questions ?

Some publications

Ibekwe-SanJuan F., Semantic metadata annotation. Tagging Medline abstracts for enhanced information access, in "Content Archtecture. Exploiting and Managing Diverse Resources", 1st Biennal Conference of the British Chapter of the International Society for Knowledge Organization (ISKO-UK09), London, UK, 22-23 June, 2009.

Ibekwe-SanJuan F., Fernandez S., SanJuan E., Charton E., Annotation of Scientific Summaries for Information Retrieval, Workshop "Exploiting Semantic Annotations in Information Retrieval", in 30th European Conference on Information Retrieval (ECIR-08), Glasgow, 30th March 2008, 70-83.

Ibekwe-SanJuan F., Chen C., Pinho R., Identifying Strategic Information from Scientific Articles through Sentence Classification, 6th International Conference on Language Resources and Evaluation Conference (LREC-08), Marrakesh, Morocco, 26 May -1st June, 2008, 5p.

Ibekwe-SanJuan F. (2005), Annotation d'indices de nouveautés dans les écrits scientifiques et techniques, Colloque "Indice, Index, Indexation, 3-4 novembre 2005, Université de Lille 3, France, 12p.

Categorization by linguistic cues

{S}19130928.
{S} Notch signaling is involved in cell fate determination along with the development of the immune system.
{S} However, very little is known about the role for Notch signaling in mast cells.
{S} **[xOBJECTIVEx]** *We investigated* the role of Notch signaling in mast cell functions.
{S} After mouse bone marrow-derived mast cells (BMMCs) or peritoneal mast cells (PMCs) were cocultured with mouse Notch ligand-expressing chinese hamster ovary cells for 5 days, we examined the mast cell surface expressions of MHC-II molecules and OX40 ligand (OX40L), Fc epsilon RI-mediated cytokine production, and the effects of the mast cells on proliferation and differentiation of naive CD4(+) T cells in vitro.
{S} **[xRESULTx]** *We showed that* BMMCs and PMCs constitutively expressed Notch1 and Notch2 proteins on the cell surface.
{S} **[xRESULTx]** *We also found* that Delta-like 1 (Dll1)/Notch signaling induced the expression of MHC-II and upregulated the expression level of OX40L on the surface of the mast cells.
{S} Dll1/Notch signaling augmented Fc epsilon RI-mediated IL-4, IL-6, IL-13, and TNF production by BMMCs.
{S} Dll1-stimulated MHC-II(+)OX40L(high) BMMCs promoted proliferation of naive CD4(+) T cells and their differentiation into T(H)2 cells producing IL-4, IL-5, IL-10, and IL-13.
{S} Dll1/Notch signaling confers the functions as an antigen-presenting cell on mast cells, which preferentially induce the differentiation of T(H)2.

The original Medline abstract

19130928. Notch signaling confers antigen-presenting cell functions on mast cells.

BACKGROUND: Notch signaling is involved in cell fate determination along with the development of the immune system. However, very little is known about the role for Notch signaling in mast cells.

OBJECTIVE: We investigated the role of Notch signaling in mast cell functions.

METHODS: After mouse bone marrow-derived mast cells (BMMCs) or peritoneal mast cells (PMCs) were cocultured with mouse Notch ligand-expressing chinese hamster ovary cells for 5 days, we examined the mast cell surface expressions of MHC-II molecules and OX40 ligand (OX40L), Fc epsilon RI-mediated cytokine production, and the effects of the mast cells on proliferation and differentiation of naive CD4(+) T cells in vitro.

RESULTS: We showed that BMMCs and PMCs constitutively expressed Notch1 and Notch2 proteins on the cell surface. We also found that Delta-like 1 (Dll1)/Notch signaling induced the expression of MHC-II and upregulated the expression level of OX40L on the surface of the mast cells. Dll1/Notch signaling augmented Fc epsilon RI-mediated IL-4, IL-6, IL-13, and TNF production by BMMCs. Dll1-stimulated MHC-II(+)OX40L(high) BMMCs promoted proliferation of naive CD4(+) T cells and their differentiation into T(H)2 cells producing IL-4, IL-5, IL-10, and IL-13.

CONCLUSION: Dll1/Notch signaling confers the functions as an antigen-presenting cell on mast cells, which preferentially induce the differentiation of T(H)2.