HELSINGIN YLIOPISTO

Structure patterns in Information Extraction

Gaël Lejeune, Research Assistant

University of Helsinki

Outline

- Overview of Information extraction
- PULS system
- French version
- Results
- Conclusion

Overview of Information Extraction

- Problem (related to semantic web) :

  Most documents are made to be readable by humans not by machines.

- Solution:

  Processing a large quantity of documents automatically and extract relevant information.

- Basical process:

  From unstructured documents no metadata

  To structured information databases

Classical approaches

Giving up the ”bag of words” concept
but keeping word granularity

- Lexical normalization *morphemes*
- Morphological analysis *words/lexical items*
- Syntactic analysis *chunks*
- Semantic analysis *phrases/sentences*
- Semantic interpretation *“meaning”*
- Discourse analysis *coreference*

Classical applications

- Business

Jouko Seppä, the head of ICL E-Business Division, has been appointed managing director…
[Person] [Old Position] has been appointed [New position]

- Terrorism

Unidentified individuals planted a bomb in front of a Mormon Church
[Perpetrator] planted a bomb [Target]

- Epidemic

4,500 people in 29 countries have been confirmed to have been infected with swine flu
[Victim] [Location] have been infected [Disease]

PULS

- MedISys provides documents to PULS
- PULS extracts events and adds interaction:
  - between documents
  - between events
- PULS provides an online database

PULS

Confident events [Events] [Advanced query] [Groups]

[Database list]

[Reset page][Help]

[Login][Chart][Map]

	Published	Source	Disease	Country	Begin	End	Total	†	Descriptor
[9] +	2009.05.08	allafrica	Cholera	Republic of Congo	2009.04	2009.04	130		130 Cholera Cases
	2009.05.08	googlenewshealth	Influenza	United Arab Emirates	_	2009.05.08	1 124		1,124 cases
	2009.05.08	idsociety	Influenza A virus H ...	USA	_	2009.05.08	2	†	two confirmed deaths
[927] +	2009.05.08	bbc	Swine Flu	UK	2009.05.08	2009.05.08	2		Two more probable sw...
	2009.05.08	alertnet	Influenza A virus H ...	Mexico	_	2009.05.08	1 112		1,112 laboratory-con...
[21] +	2009.05.08	HonululuAdvetrtiser	Swine Flu	USA/Hawaii	2009.05.07	2009.05.07	3		The first three case...
[21] +	2009.05.08	HonululuAdvetrtiser	Swine Flu	USA/Hawaii	2009.05.07	2009.05.07	--		--
[19] +	2009.05.08	HonululuAdvetrtiser	Swine Flu	Japan	2009.05.07	2009.05.07	0		no confirmed cases
[58] +	2009.05.08	HonululuAdvetrtiser	Swine Flu	USA/Illinois	2009.05.07	2009.05.07	1		A Japanese boy
[218] +	2009.05.08	HonululuAdvetrtiser	Influenza	worldwide	2009.05.07	2009.05.07	--		people
[142] +	2009.05.08	moreoverhealth	Swine Flu	worldwide	2009.05.07	2009.05.07	--		people
[2] +	2009.05.08	eastandard	Dysentery	Kenya	2009.05.07	2009.05.07	37		37 dysentery cases
[65] +	2009.05.08	googlenewshealth	Swine Flu	USA/Florida	2009.05.07	2009.05.07	22		22 probable cases
[74] +	2009.05.08	yle_en	Influenza A virus H ...	Hong Kong	2009.05.01	2009.05.01	--		resident
[5341] +	2009.05.08	guardian	Influenza	Mexico	2009.05.08	2009.05.08	42	†	42 people
[927] +	2009.05.08	guardian	Swine Flu	UK	2009.05.07	2009.05.07	1		confirmed cases
[927] +	2009.05.08	guardian	Swine Flu	UK	2009.05.07	2009.05.07	34		confirmed cases
[61] +	2009.05.08	hpa	Swine Flu	Europe	2009.05.08	2009.05.08	--		humans
[61] +	2009.05.08	hpa	Swine Flu	Europe	2009.05.08	2009.05.08	--		humans
[2] +	2009.05.08	dailymail	Meningitis	UK	2009.05.08	2009.05.08	1	†	A 17-year-old girl

Guideline

Type of event	Explanation/guidelines	Relevance score
highly relevant	new information	1
quite relevant	important update, on-going developments	2
less relevant	current events, but this is a review article	3
low relevance	historical but potentially useful as background information	4
not relevant	non-specific events, non-factual, article focusing on secondary topics	5
UNCLEAR	WRONG EVENT or wrong type of event	-1

Multilingual goal

- One language is not sufficient
- Machine translation is not ready to help us
- We have some constraints:
  - Resources are hard to build
  - More steps you have, more errors you may get

PULS French System

- Two fields of linguistics are almost ignored:
  - Stylistics
  - Pragmatics

- Though they give us two useful ”tools”:
  - 5W rule
  - Pertinence/effectiveness rule

5W rule

- Main information is in the top of the document, for our purpose it will be:
  - What: Disease
  - Where: Country
  - Who: Cases
  - When: Date

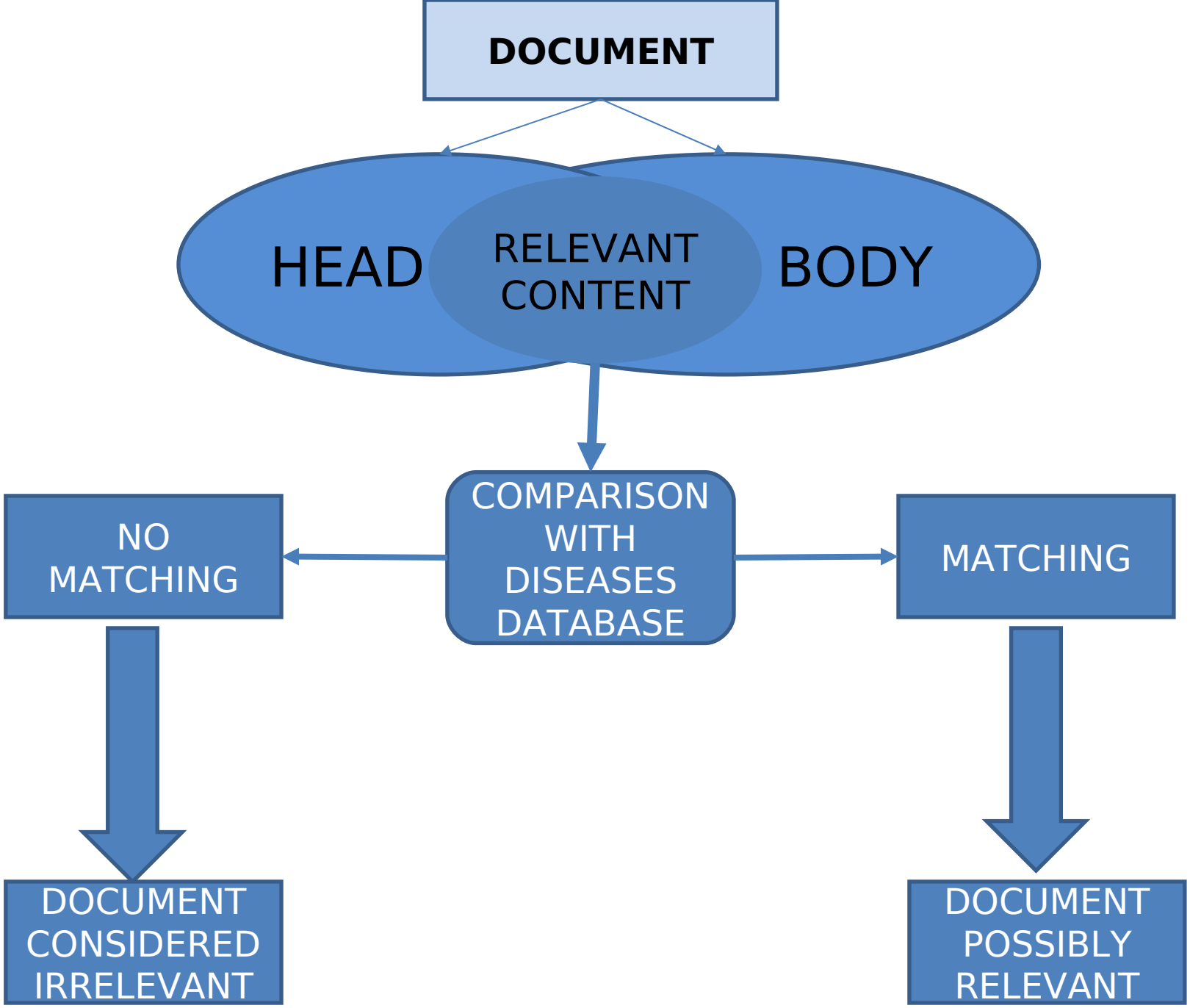
Pertinence rule

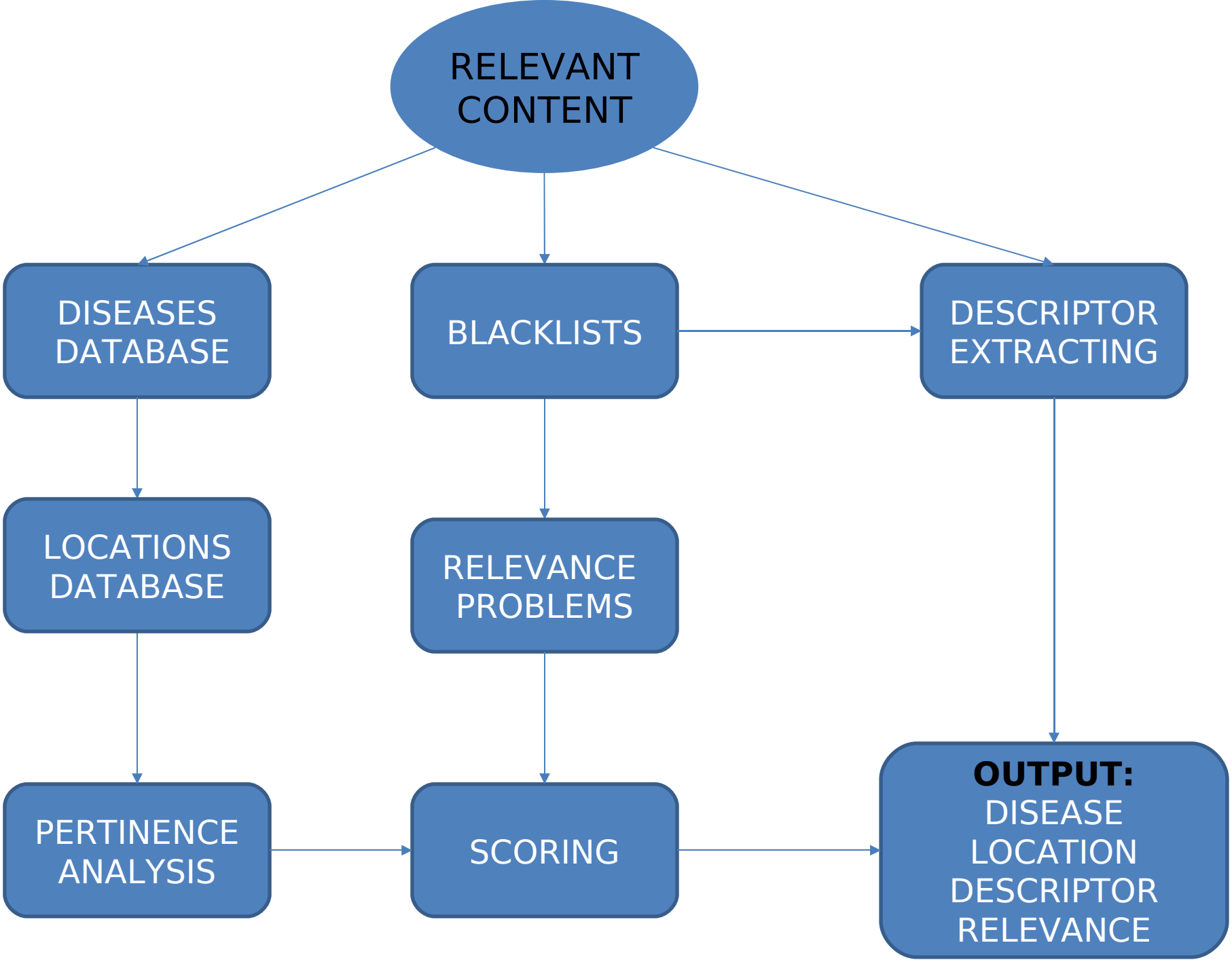
- One important information= one article
  - If you have two events, one is less important
  - The most important is the first to be related

- Important matters are related explicitely
  - The headline is decisive
  - All that can be ambiguous is explicated

Components

- Disease database: 150 items
- Location database: 400 items
- Blacklists: 20 items


DOCUMENT
HEAD
RELEVANT CONTENT
BODY
COMPARISON WITH DISEASES DATABASE
NO MATCHING
MATCHING
DOCUMENT CONSIDERED IRRELEVANT
DOCUMENT POSSIBLY RELEVANT


RELEVANT CONTENT
DISEASES DATABASE
BLACKLISTS
DESCRIPTOR EXTRACTING
LOCATIONS DATABASE
RELEVANCE PROBLEMS
PERTINENCE ANALYSIS
SCORING
OUTPUT:
DISEASE
LOCATION
DESCRIPTOR
RELEVANCE

Example

Le **choléra** peut affecter **60.000 personnes** (Pana) –

L'épidémie de **choléra** qui fait rage au Zimbabwe pourrait affecter **60.000 personnes** si elle n'est pas maîtrisée de toute urgence, a prévenu, hier vendredi, l'Organisation mondiale de la santé (OMS), dans un communiqué rendu public à son siège à Genève, en Suisse, rejetant les déclarations....

L'organisation onusienne note que beaucoup de gens dans ce pays continuent encore dutiliser de leau non potable et de vivre dans des conditions peu hygiéniques, ce qui est à la base de cette épidémie. LOMS a dépêché une équipe dexperts au Zimbabwe pour aider ce pays à lutter contre l'épidémie de **choléra**, la pire qui frappe ce pays dAfrique australe peuplé de 14 millions d'habitants. Le président Zimbabwéen, Robert mugabe, a déclaré jeudi que la maladie a été enrayée, une affirmation démentie par lOMS. Lorganisation onusienne note que lépidémie de **choléra** continue de plus belle au Zimbabwe et quelle pourrait affecter près de **60.000 personnes**. Jusquici, plus de **600 personnes sont mortes** de la maladie et près de **20.000 autres ont été infectées**....

DISEASE **LOCATION** **CASES**

Results

- Corpus of 1200 documents containing 210 manually tagged as relevant

	Corpus	Extracted	
Event	210	196	93% Recall
No event:	990	28	86% Precision
Total	1200	224	89% F-Measure

- **Locations** extracted:
  86% good unique disease/location pairs

- **Cases found**:
  93% of descriptors are good

On-going work on Spanish

- Same components as French version:
  - “Easy to build” databases
  - Keeping the same scripts

- Test on a corpus of 100 documents:
  - Recall 71%
  - Precision 80%
  - All documents had good descriptors

По русски

- «**Свиной грипп**» шествует по **миру**: уже 4379 заболевших в 29 странах
- Опубликована: 10 мая 2009 19:53:11 По данным Всемирной организации здравоохранения количество заболевания **гриппом** A/H1N1 увеличилось до **4379** в 29 странах **мира**.
- Еще в субботу ВОЗ сообщал, что количество заболевших **3440 человек**. НА сегодняшний момент 45 человек уже умерло от «**свиного гриппа**» в Мексике, 2 – в США, 1 – в Канаде, 1 – в Коста-Рике: итого – 49 человек. По прежнему, большинство заболевших Мексике и США, зарегистрированы случаи вируса в Латинской Америке, Европе и Азии. ВОЗ призывает людей с ослабленным иммунитетом отложить поездки в другие страны и сразу же обращаться к врачу при появлении первых симптомов **гриппа**. Напомним, что эпидемия вызвана мутировавшим вирусом **гриппа** типа А. Симптомы – повышенная температура, головная и мышечная боль, иногда рвота и диарея. Уровень угрозы пандемии по шестибальной шкале по-прежнему равен 5. Ранее ученые неоднократно заявляли, что нынешняя эпидемия **гриппа** вряд ли повторит "испанку", которая в 1918-1920 годах унесла более 20 миллионов жизней, поскольку теперь медики и эпидемиологи намного больше знают о возбудителе **гриппа** и механизмах распространения болезни, сообщает РИА Новости.

DISEASE **LOCATION** **CASES**

Conclusion

- The promising scores we got from that experimental try has convinced us that there are important improvements to get from “text granularity rules”.

- Our next step will be to test our system on other Romance languages (for instance Italian) then to other Indo-European ones.

- If we can keep the idea and the simplicity of it in that number of languages we would be able to say that we can monitor confidently an important part of the epidemic data in the world.

HELSINGIN YLIOPISTO

Thank you for listening

Спасибо болшой