Text Mining, Information and Fact Extraction
Part 4: Applications

Marie-Francine Moens
Department of Computer Science
Katholieke Universiteit Leuven, Belgium
sien.moens@cs.kuleuven.be

General setting

- Information extraction: has received during decades a large interest because of its applicability to many types of information
- In IR context: interest in IE from text is boosted by growing interest in IE in other media (e.g., images, audio)

- Note: performance statistics given in this chapter are only indicative and refer to a particular setting (corpus, features used,classification algorithm, ...)

Overview

- Generic versus domain-specific character of IE tasks
- Possible applications:
  - Processing of news texts
  - Processing of biomedical texts
  - Intelligence gathering
  - Processing of business texts
  - Processing of law texts
  - Processing of informal texts

Overview

- Specific case studies:
  - Recognizing emotions expressed towards product or person (joint work with Erik Boiy)
  - Recognizing actions and emotions performed or expressed by persons (joint work with Koen Deschacht)

Generic versus domain specific character

- Generic information extraction and text mining: use of generic ontology or classification scheme
  - Named entity recognition (person, location names, ...)
  - Noun phrase coreference resolution
  - Semantic frames and roles, ...
- Domain-specific information extraction and text mining: use of ontology of domain-specific semantic labels
- Techniques and algorithms are fairly generic

Processing news texts

- Very traditional IE boosted by Message Understanding Conferences (MUC) in late 1980s and 1990s (DARPA), followed by Automatic Content Extraction (ACE) initiative and Text Analysis Competition (TAC) (NIST)
- Tasks:
  - Named entity recognition
  - Noun phrase coference resolution
  - Entity relation recognition
  - Event recognition (who, what, where, when)

Nastia Liukin wins women's gymnastics all-around gold

Adjust font size:

Nastia Liukin of the United States edged her compatriot Shawn Johnson to win the women's all-around after a breathtaking Olympic gymnastics competition on Friday.

WHO?

Finishing one-two on the podium, the American duo let the hosts' gold rush in the gymnastics pause after China wrapped up all the first three titles (men's team, women's team, men's all-round) in the National Indoor Stadium.

WHEN?

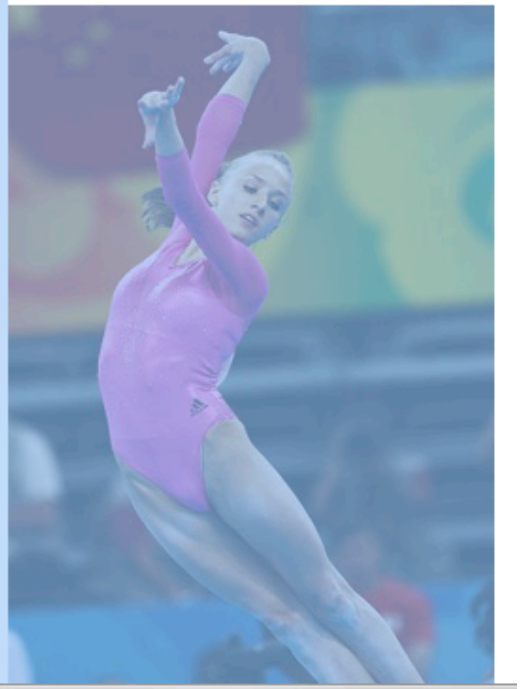
Liukin collected 63.325 points, after flawless exercises on each of the four apparatus with a good combination of difficulty and quality, beating Johnson by 0.600. The bronze medal went to Yang Yilin of China in 62.650.

WHAT?

Coming out in her first Olympic Games, the 18-year-old Liukin struck the most coveted gold medal which she had waited for years.

WHERE?

She was an unlucky runner-up in the 2005 world championship in Melbourne, beaten by Chellsie Memmel by 0.001 points in her debut to the international arena, and the title also eluded her in the following two championships.



www.china.org.cn

Processing news texts

- **Named entity recognition**:
  - Person, location, organization names
  - Mostly supervised: Maxent, HMM, CRF
  - Approaches human performance: in literature sometimes above 95% F_1 measure

[Bikel et al. ML1999] [Finkel et al. 2006]

- **Noun phrase coreference resolution**:
  - Although unsupervised (clustering), and semi-supervised (co-training), best results with supervised learning: F_1 measures of 70% and more are difficult to reach; also kernel methods

[Ng & Cardie ACL 2002] [Ng & Cardie HLT 2003] [Versley et al. COLING 2008]

Processing news texts

- **Entity relation recognition**:
  - use of supervised methods: e.g., kernel methods: F_1 measures fluctuate dependent on number of training examples and difficulty of the relational class (ambiguity of the features)

[Culotta & Sorensen ACL 2004] [Girju et al. CSL 2005]

- **Event recognition**:
  - in addition: recognition and resolution of:
    - temporal expressions: TimeML
    - spatial expressions: FrameNet and Propbank

[Pustejovsky et al. IWCS-5 2003] [Baker et al. COLING-ACL 1998]
[Morarescu IJCAI 2007] [Palmer et al. CL 2005]

Processing news texts

- Challenges:
  - Cross-document, cross-language and cross-media (video !):
    - named entity recognition and resolution
    - event recognition:
      - including cross document temporal and spatial resolution

Processing biomedical texts

- Many ontologies or classification schemes and annotated databases are available:
  - E.g., Kyoto Encyclopedia of Genes and Genomes, Gene Ontology, GENIA dataset
- Tasks:
  - Named entity recognition
  - Relation recognition
  - Location detection and resolution

A UNIQUE CONTRIBUTION OF HEAT SHOCK TRANSCIPTION FACTOR 4 IN OCULAR LENS DEVELOPMENT AND FIBER CELL DIFFERENTIATION

Jin-Na Min, Yan Zhang, Demetrius Moskophidis, and Nahid F. Mivechi*

Institute of Molecular Medicine and Genetics, Molecular Chaperone Biology/Radiobiology Program, Medical College of Georgia, Augusta, GA, 30912.
*mivechi@immag.mcg.edu

INTRODUCTION Defects in the development and physiology of the eye lens as a result of gene mutations can cause cataracts, the commonest form of visual impairment in humans. Congenital cataracts account for around 10% of cases of childhood blindness, one-half of which have a genetic cause [1]. Ocular lens development is coordinated by expression of growth and transcription factors such as Pax6, FoxE3, Six3, Prox1, Sox2/3, Maf, Pitx3, AP-2a. Normally after formation of the lens vesicle, which is filled by elongated cells on its posterior surface (primary fibers), mitotically active cells from the monolayer of cuboidal epithelial cells at the anterior lens pole travel towards the equator where they elongate and differentiate into secondary lens fibers. Maturation of fiber cells is accompanied by

Processing biomedical texts

- **Named entity recognition**: difficult:
  - boundary detection:
    - capitalization patterns: often misleading
    - many premodifiers or postmodifiers that are part or not of the entity (**91 kDA protein**, activated **B cell** lines)
  - polysemous acronyms and terms: e.g., **PA** can stand for **pseudomonas aeruginosa**, **pathology** and **pulmonary artery**
  - synonymous acronyms and terms
- Supervised context dependent classification: HMM, CRF: often F_1 measure between 65-85%

[Zhang et al. BI 2004]

Processing biomedical texts

- **Entity relation recognition**:
  - Protein relation extraction
  - Literature based gene expression analysis
  - Determination of protein subcellular locations
  - Pathway prediction (cf. event detection)
    - methods relying on symbolic handcrafted rules, supervised (e.g., CRF) and unsupervised learning

[Stapley et al. PSBC 2002] [Glenisson et al. SIGKDD explorations 2003] [Friedman et al. BI 2001] [Huang et al. BI 2004] [Gaizauskas et al. ICNLP workshop 2000]

Intelligence gathering

- Evidence extraction and link discovery by police and intelligence forces from narrative reports, e-mails and other e-messages, Web pages, ...
- Tasks:
  - Named entity recognition, but also brands of cars, weapons
  - Noun phrase coreference resolution, including strange aliases
  - Entity attribute recognition
  - Entity relation recognition
  - Event recognition (recognition and resolution of temporal and spatial information; frequency information !)

www.kansascitypi.com

Intelligence gathering

- See above news processing
- **Entity attribute recognition**: often visual attributes, very little research;
  - recognition of visual attributes in text based on association techniques (e.g., chi square) of word and textual description of image

African violets (Saintpaulia ionantha) are small, flowering houseplants or greenhouse plants belonging to the Gesneriaceae family. They are perhaps the most popular and most widely grown houseplant. Their thick, fuzzy leaves and abundant blooms in soft tones of violet, purple, pink, and white make them very attractive…

A small girl looks up at a person dressed in the costume of an anima which could be "Woody Woodchuck" at the State Fair in Salem, Oregon.

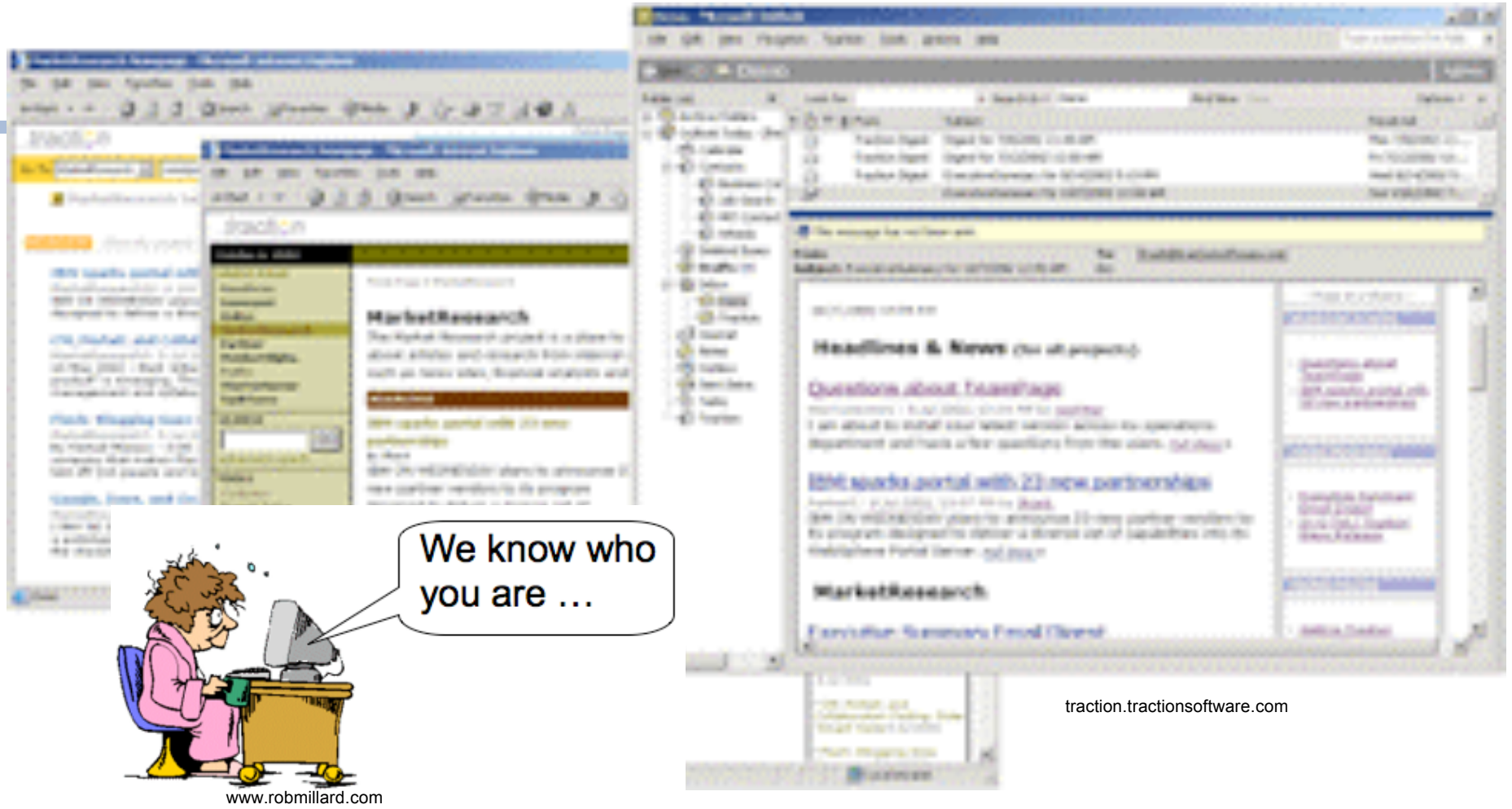
[Boiy et al.TIR 2008]

Intelligence gathering

- Challenges:
  - Texts are not always well-formed (spelling and grammatical errors): drop in F_1 measures compared to standard language
  - Often important to detect the single instance
  - Combination with mining of other media (e.g., images, video)
  - Recognition of temporal and spatial relationships, recognition of other rhetorical relationships (e.g., causal) [Hovy AI 1993] [Mann & Thompson TR 1997] [Mani 2000]
  - Extracted information is often used to build social networks, which can be mined for interesting patterns

Processing business texts

- Wealth of information can be found in technical documentation, product descriptions, contracts, patents, Web pages, financial and economical news, blogs and consumer discussions
- Business intelligence (including competitive intelligence): mining of the above texts



www.robmillard.com

traction.tractionsoftware.com

Processing business texts

- Tasks:
  - Named entity recognition: including product brands
  - Entity attributes: e.g., prices, properties
  - Sentiment analysis and opinion mining

Processing law texts

- Processing legislation, court decisions and legal doctrine
- Tasks:
  - Named entity recognition
  - Noun phrase coreferent resolution
  - Recognition of factors and issues
  - Recognition of arguments
  - Link mining
- For a long time: low interest, but since 2007: TREC legal track (NIST)

UNITED NATIONS CONVENTION ON CONTRACTS FOR THE INTERNATIONAL SALE OF GOODS (1980) [CISG]

For U.S. citation purposes, the UN-certified English text is published in 52 Federal Register 6262, 6264-6280 (March 2, 1987); United States Code Annotated, Title 15, Appendix (Supp. 1987).

Linked Table of Treaty Sections

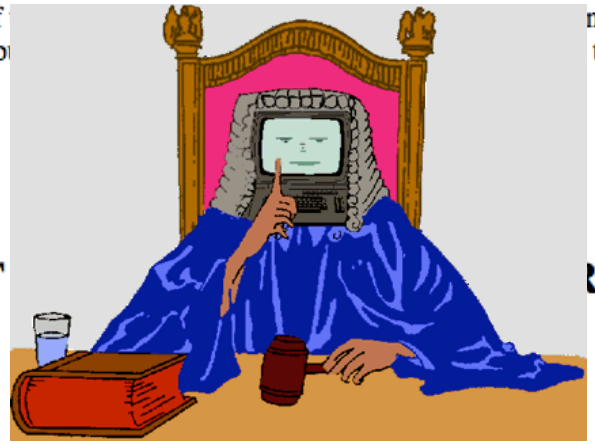
THE STATES PARTIES TO THIS CONVENTION,

BEARING IN MIND the broad objectives in the resolutions adopted by the sixth special session of the General Assembly of the United Nations on the establishment of a New International Economic Order,

CONSIDERING that the development of international trade on the basis of equality and mutual benefit is an important element in promoting friendly relations among States,

BEING OF THE OPINION that the adoption of national sale of goods and take into account the different social, economic and legal systems would contrib trade and promote the development of international trade,

HAVE AGREED as follows:



SPHERE OF RAL PROVISIONS

Processing law texts

- **Recognition of factors and issues in cases**:
  - factor = a certain constellation of facts
  - issue = a certain constellation of factors
- Limited attempts to learn factor patterns from annotated examples based on a naive Bayes and decision tree learners
- Difficulties:
  - ordinary language combined with a typical legal vocabulary, syntax and semantics: making disambiguation, part-of-speech tagging and parsing less accurate

[Brüninghaus & Ashley 2001]

Processing law texts

- **Recognition of argumentation** and its composing arguments in cases:
  - an argument is composed of zero or more premises and a conclusion
  - discourse structure analysis
- Difficulties:
  - see recognition of factors and issues
  - discourse markers are ambiguous or absent
  - argument are nested (conclusion of one argument is premise of another argument)
  - difficult style: humans have difficulty to understand the content

[Mochales Palau & Moens 2008]

Processing informal texts

- Many texts diverge from standard language when created or when processed:
  - Spam mail
  - Blog texts
  - Instant messages
  - Transcribed speech
  - ...

operator: thanks for calling the ibm customer service center this is john am i speaking with mark
caller: yes
operator: hey mark what's going on
caller: well i'm trying to connect to the at and t net client and i got an error it came back and gave me error one twenty it says invalid access list and i'm not sure what it means by that
operator: one twenty invalid access list
caller: uhhuh
operator: let's go and take a look at your setup real quick and see see what's going on there
caller: ok
...
caller: well it ah trying to ah oh i got something um the t and t at and t net client got A GRAPHIC ON MY SCREEN so it's connecting and it's counting the connect time so let's see it's got a little downer i think i can minimise it by clicking on it oops ah that's not it ah yeah i can minimise it ok great and i'll just try to get into lotus notes and see what happens

operator: you for calling i b m customer service center this is john to bookmark
caller: yes
operator: hey mark what's going on
caller: well i'm trying to connect to the uh a t n t net client and uh i got an error or came back and gave me an error one twenty it says invalid access list and i'm not sure what it means by that
operator: ok one twenty invalid access list
caller: ok
operator: it's gonna take a look at your setup real quick and see what see what's going on there
caller: ok
...
caller: i with to trying to install ok i've got something uhhuh at a and t a t n t net client HAVE GLASSES ON MY SCREEN fast so it's connecting and it's counting the connect time off so let's see if that'll down there at the ticket number a second notes left that's not it yeah i can minimize ok great and i'll just try to get into uh lotus notes and see what happened

Figure 2: Transcripts of a excerpt of a call. Left - a manual transcript. Right - an automatic 1-best path transcript of the same call, with 36.75% WER level .

[Mamou et al. SIGIR 2006]

Processing informal texts

- Accuracy of the extraction usually drops proportional with the amount of noise
- Solutions:
  - Preprocessing: e.g., most likely normalization based on string edit distances, language models
  - Incorporating different hypotheses into the extraction process

Processing informal texts

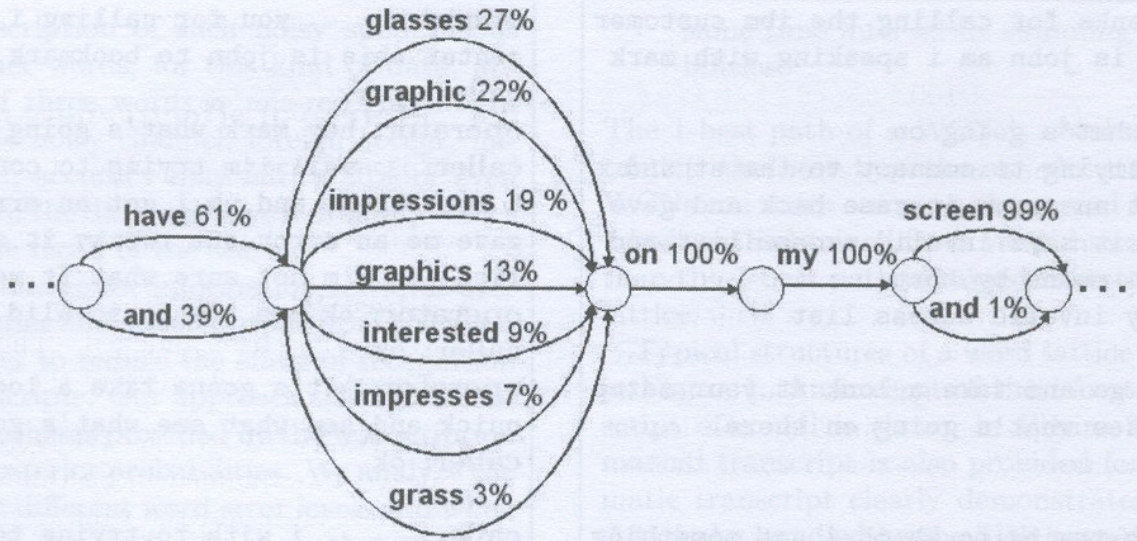


Figure 3: A fragment of the WCN of the call appeared in Figure 2. Timestamps have been omitted for clarity.

[Mamou et al. SIGIR 2006]

Case studies

Case 1: Emotion expressed towards person or product

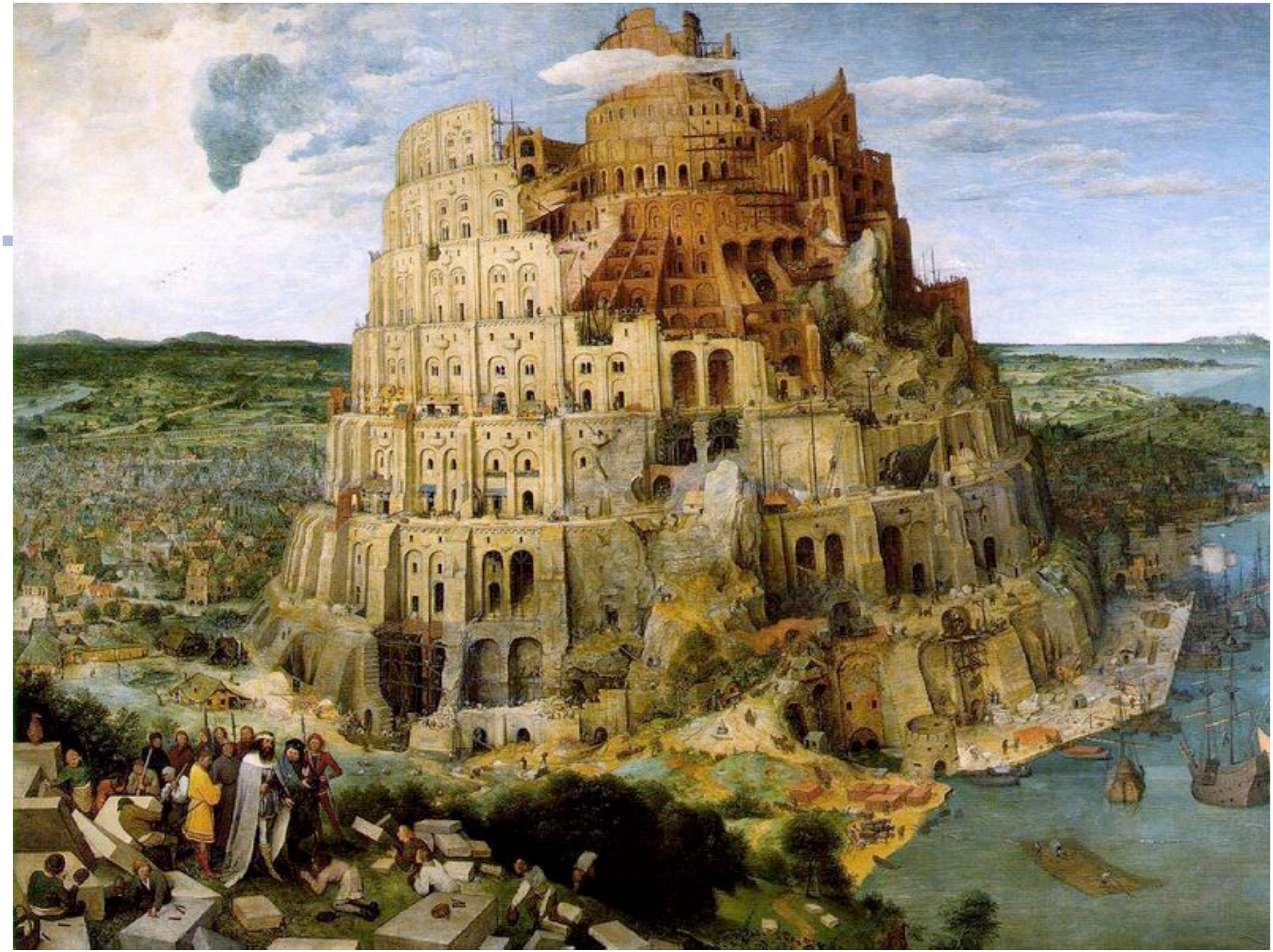
- Learning emotion patterns in **blog, review and news fora** texts:
  - Positive, negative and neutral feeling

- Problems:
  - Large variety of expressions (noisy texts !!!) and relatively few annotated examples
  - Emotion is attributed to an entity
  - Language/domain portability (English, Dutch and French blogs)
  - How to reduce the annotation of training examples?

\- *The movie really seems to be spilling the beans on a lot of stuff we didnt think we hand if this is their warm up, what is going to get us frothing in December*

-+ *de grote merken mogen er dan patserig uitzien en massa's pk hebben maarals de bomen wat dicht bij elkaar staan en de paadjes steil enbochtig,dan verkies ik mijn Jimny .*

\+ *L'é tro bel cet voitur Voici tt ce ki me pasione ds ma petite vi!!!é tt mé pote é pl1 dotre truk!!!Avou de Dcouvrir*

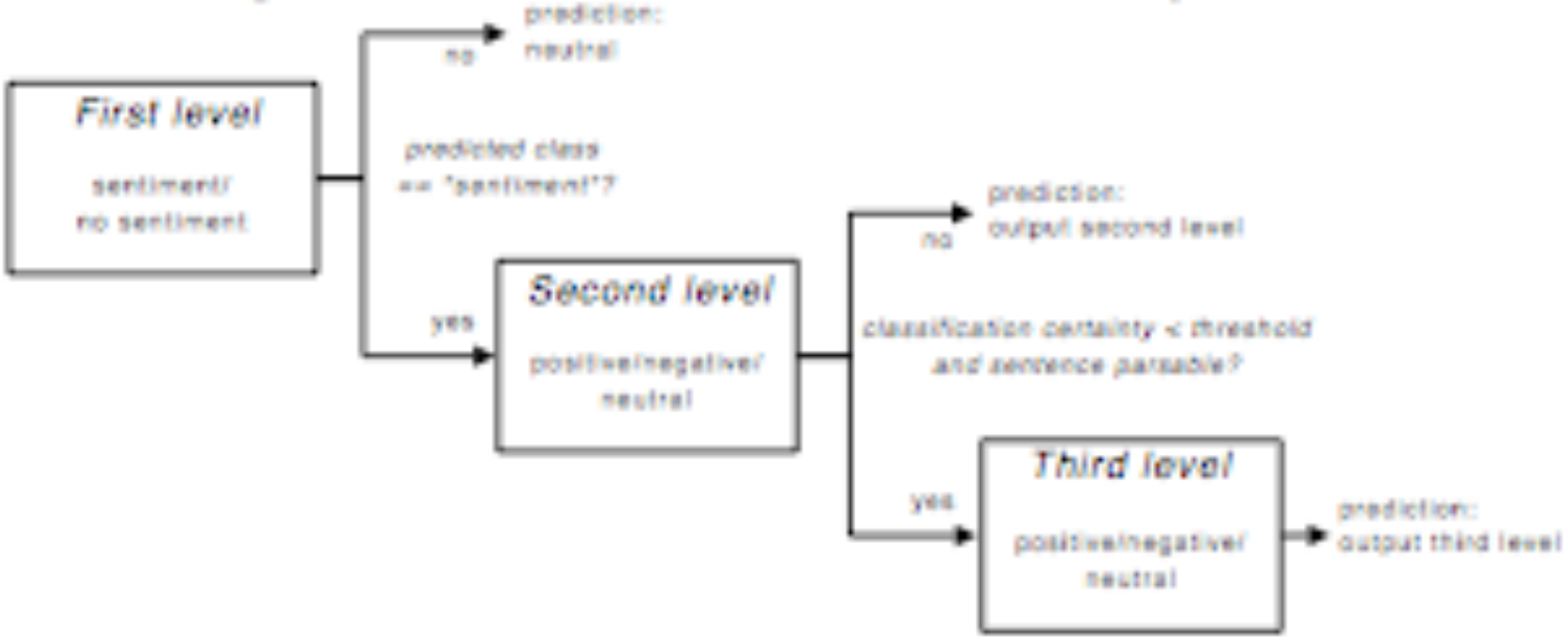


Case 1: Emotion is expressed towards person or object

- Solutions tested:
  - Feature extraction
  - Single classifier versus a cascaded classifier versus bagged classifiers
  - Active learning

[Boiy & Moens IR 2008]

Figure 1: Instantiation of the cascade model used in our experiments.



[Boiy & Moens IR 2008]

Case 1: Emotion is expressed towards person or object

- Corpus:
  - blogs: e.g., skyrock.com, lifejournal.com, xanga.cpm, blogspot.com; review sites: e.g., amazon.fr, ciao.fr, kieskeurig.nl; news fora: e.g., fok.nl, forums.automotive.com
  - 750 positive, 750 negative and 2500 neutral sentences for each language
  - interannotator agreement: κ = 82%
- Codes in the table below:
  - SC uni: unigram features
  - SC uni-lang: + language (negation, discourse) features
  - SC uni-lan-dist: + distance in number of words with entity feature

Table 2: Our best results in terms of accuracy, precision, recall and F-measure (F_1) using the English (a), Dutch (b) and French (c) corpora. For English, Dutch and French we implemented respectively an MNB, an SVM and an ME classifier – 10 fold cross-validation.

(a) English

Architecture	Accuracy	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	83.30	69.09/85.48/85.93	55.73/82.40/91.84	61.70/83.91/88.79
Cascade with layers 1 and 2	83.10	70.49/87.72/84.61	54.13/79.07/93.00	61.24/83.17/88.61
SC uni-lang	83.03	69.59/86.77/85.08	56.13/79.60/92.12	62.14/83.03/88.46
SC uni-lang-dist	80.23	60.59/78.78/86.57	59.87/82.67/85.60	60.23/80.68/86.08
SC uni	82.73	68.01/85.63/85.53	58.40/78.67/91.24	62.84/82.00/88.29

(b) Dutch

Architecture	Accuracy	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1,2 and 3	69.03	63.51/53.30/72.20	42.93/31.20/88.20	51.23/39.36/79.40
Cascade with layers 1 and 2	69.80	66.60/58.31/71.66	41.73/29.47/90.32	51.31/39.15/79.92
SC uni-lang	69.05	60.39/52.59/73.63	49.60/33.87/85.44	54.47/41.20/79.10
SC uni-lang-dist	68.85	61.08/54.52/72.20	43.73/30.53/87.88	50.97/39.15/79.27
SC uni	68.18	58.73/49.58/73.24	48.00/31.73/85.16	52.82/38.70/78.75

(c) French

Architecture	Accuracy	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	67.68	50.74/55.88/71.90	27.47/38.67/88.44	35.64/45.71/79.32
Cascade with layers 1 and 2	67.47	52.69/53.96/71.56	26.13/38.13/88.68	34.94/44.69/79.21
SC uni-lang	65.97	47.67/50.33/72.18	30.00/40.67/84.36	36.82/44.99/77.79
SC uni-lang-dist	65.97	47.67/50.33/72.18	30.00/40.67/84.36	36.82/44.99/77.79
SC uni	65.83	45.67/50.82/72.23	28.80/41.33/84.28	35.32/45.59/77.79

[Boiy & Moens IR 2008]

Inter-annotator agreement

- **Kappa statistic**: agreement rate when creating 'gold standard' or 'ground truth' corrected for the rate of chance agreement

  where $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

  $P(A)$= proportion of the annotations on which the annotators agree

  $P(E)$ = proportion of the annotations on which annotations would agree by chance

- $\kappa > 0.8$: good agreement
- $0.67 <= \kappa <= 0.8$: fair agreement
- More than 2 judges: compute average pairwise κ

Active learning

- **Active learning** = all examples to train from are labeled by a human, but the set of examples is carefully selected by the machine
- (Starts with labeled set on which the classifier is trained)
- Repeat
  - 1 or a bucket of examples are selected to label:
    - which are classified by the current classifier as most uncertain (informative examples)
    - that are representative or diverse (e.g., found by clustering)
- Until the trained classifier reaches a certain level of accuracy on a test set

Active learning

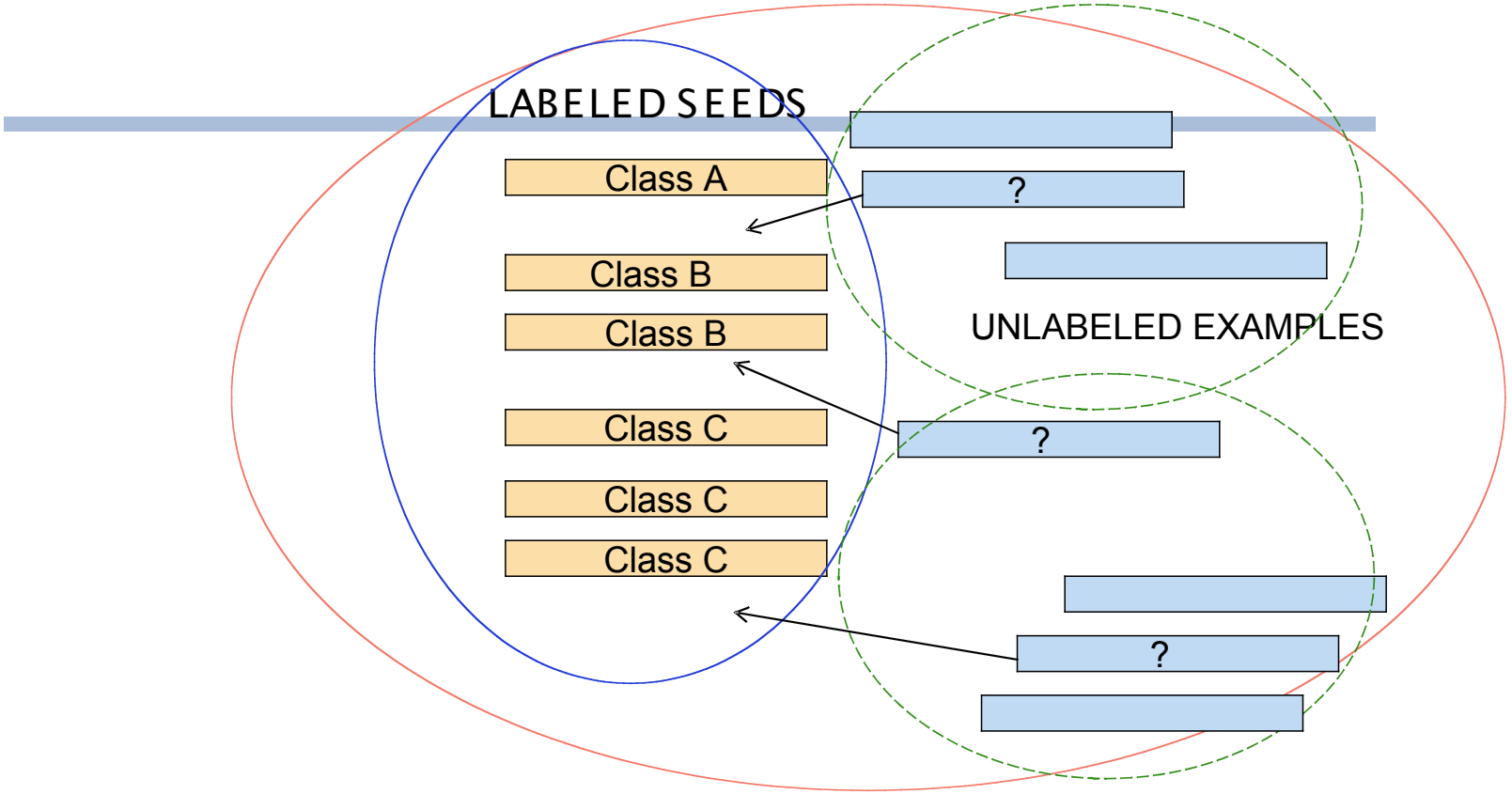


Fig. 6.5. Active learning: Representative and diverse examples to be labeled by humans are selected based on clustering.

Case 1: Emotion is expressed towards person or object

- Active learning techniques tested on English corpus:
  - Uncertainty sampling (US): to find informative examples
  - Relevance sampling (RS): to find more negative examples
  - Combination of US and RS yielded best results:

Table 11: Comparison of RS and US for the MNB uncertainty sampling method using seed size 150 and batch size 10. The number after ± is the standard deviation – averaged over 5 runs.

	Accuracy		F-measure pos		F-measure neg	
#Ex	RS	US	RS	US	RS	US
150	68.10±00.39	68.10±00.39	35.05±06.70	35.05±06.70	26.64±03.21	26.64±03.21
200	73.45±01.01	70.23±00.60	36.50±08.75	33.74±08.32	30.97±02.32	27.67±03.06
250	75.88±01.20	74.25±01.36	37.41±09.15	35.02±07.98	33.43±01.40	31.46±03.55
300	77.53±00.88	76.74±01.61	36.96±10.48	37.91±02.95	33.65±02.75	33.20±04.99
350	78.40±01.06	77.79±01.46	38.63±09.60	40.51±03.10	31.30±06.08	34.47±07.12
400	78.46±00.71	78.25±01.59	38.26±10.33	41.06±02.17	30.52±06.44	34.38±06.30
450	79.21±00.98	79.42±01.27	39.30±06.95	42.08±03.98	31.87±05.94	36.62±05.24
500	79.54±00.70	80.06±01.04	40.15±06.19	44.40±03.63	33.30±05.40	38.21±05.97

Case 2: Person performs action or expresses emotion

- **Semantic role labeling:**

Recognizing the basic event structure of a sentence ("**who**" "**does what**" "to whom/what" "when" "**where**" ...): semantic roles that form a semantic frame

Maria	**Sharapova**	**walks**	**towards**	**the**	**field**.
x_1		x_2	x_3		x_4
y_1		y_2	y_3		y_4
actor		**movementAction**	**toLocation**		**toLocation**

CLASS (EU: 2006-2008)

Source: Buffy

Text of script: 51: Shot of Buffy opening the refrigerator and taking out a carton of milk.

Willow hugs Buffy.

- **Semantic role and frame** detection:
  - Supervised learning (state of the art)

  [Gildea & Jarowsky CompLing 2002][CompLing 2008]
  - Our task:
    - weakly supervised learning
    - combine with evidence from the images (e.g., movement)

Case 2: Person performs action or expresses emotion

- Classification of semantic frames in text: validation of 353 sentences (1 episode) from transcripts of fans of "Buffy the Vampire Slayer" (trained on 7 episodes)
- Evaluation of several classification models:
  - Supervised learning:
    - HMM
    - CRF
  - Semi-supervised: learning from unlabeled examples: learning of multiple mixture models, inference based on expectation maximization, approximate inference (Markov chain Monte Carlo sampling methods)

[Deschacht & Moens Technical Report 2008]

Case 2: Person performs action or expresses emotion

- Problem:
  - large number of patterns that signal a semantic frame/role
  - relies on sentence parse features which might be erroneous
- Results might be improved by sentence simplification techniques:
  - application of a series of hand-written rules for syntactic transformation of the sentence, where the weights of the rules and the SRL model is learned

  [Vickrey & Koller ACL 2008]

Conclusions

- Use of current information extraction technologies yield valuable input for:
  - Automatic search and linking of information
  - Automatic mining of extracted information
- But also can offer a competitive advantage for businesses:
  - Knowledge on competitors' products, prices, contacts, ...
  - Knowledge of consumers' attitudes about products, ...
  - ...
- But not always transparent what kind of information can be found, linked, inferred, ...
- So, be careful what you write ...

TIME (IWOIB: 2006-2007)

- Advanced Time-Based Text Analytics
- Partner: Attentio, Belgium

CLASS (EU FP6: 2006-2008)

- Cognitive Level Annotation Using Latent Statistical Structure
- Partners: K.U.Leuven, INRIA, Grenoble, France, University of Oxford, UK, University of Helsinki, Finland, Max-Planck Institute for Biological Cybernetics, Germany

References

Baker, C.F., Fillmore, C.J. & Lowe, J.B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, *Montreal, Canada*.

Bikel, D. M., Schwartz R. & Weischedel, R.M. (1999). An algorithm that learns what's in a name. *Machine Learning,* 34, 211-231.

Brüninghaus, S. & Ashley, K.D. (2001). Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th International Conference on Artificial Intelligence and Law* (pp. 42-51). New York: ACM.

Boiy, E. & Moens M. -F. (2008) A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval* (accepted for publication), 30 p.

Boiy, E., Deschacht, K. & Moens M.-F. (2008) Learning visual entities and their visual attributes from text corpora In *Proceedings of the 5thInternational Workshop on Text-based Information Retrieval* . IEEE Computer Society Press.

Cullota, A. & Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 424-430). East Stroudsburg, PA: ACL.

Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzetsky, A. (2001). GENIES: A natural language processing system for the extraction of molecular pathways from journal articles. *ISMB* (*Supplement of Bioinformatics*), 74-82.

Finkel, J. et al. (2005). Reporting the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics 2005,* 6 (Suppl I): S5.

References

Gaizauskas, R. J., Demetriou, G. & Humphreys, K. (2000). Term recognition and classification in biological science journal articles. In *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP* (pp. 37-44).

Glenisson, P., Mathijs, J., Moreau, Y. & De Moor, B. (2003). Meta-clustering of gene expression data and literature-extracted information. *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 5 (2), 101-112.

Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. Computational Linguistics, 28 (3), 245-288

Girju, R., Moldovan, D.I., Tatu, M. & Antohe, D. (2005). On the semantics of noun compounds. *Computer Speech and Language,* 19 (4), 479-496.

Hovy, E. (1993). Automatic discourse generation using discourse structure relations. *Artificial Intelligence,* 63 (1-2), 341-385.

Huang, M. et al. (2004). Discovering patterns to extract protein-protein interactions from full text. *Bioinformatics,* 20 (18), 3604-3612.

Mamou, J. Carmel, D. & Hoory R. (2006). Spoken document retrieval from call-center conversations. In *Proceedings of Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development of Information Retrieval* (pp. 51-58). New York: ACM.

Mann, William C. and Sandra A. Thompson (1987). *Rhetorical Structure Theory: A Theory of Text Classification*. ISI Report ISI/RS-87-190. Marina del Rey, CA: Information Sciences Institute.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: The MIT Press.

Morarescu P. (2007). A Lexicalized Ontology for Spatial Semantics. In *Proceedings of the IJCAI-2007 Workshop on Modeling and Representation in Computational Semantics*.

References

Ng, V & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 104-111). San Francisco, CA: Morgan Kaufmann.

Ng, V. & Cardie, C. (2003). Weakly supervised natural language learning without redundant views. In *Proceedings of the Human Language Technology Conference* (pp. 183-180). East Stroudsburgh, PA: ACL.

Palmer M., Gildea D., Kingsbury P. (2005). The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31 (1), 2005.

Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A. & Katz G. (2003). TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5 Fifth International Workshop on Computational Semantics*, 2003.

Stapley, BJ, Kelley LA and Sternberg MJ (2002). Predicting the sub-cellular location of proteins from using support vector machines. *Pacific Symposium Biocomputing*, 374-385.

Versley, Y., Moschitti, A., Poesio M. & Yang,, X. (2008). Coreference systems based on kernel methods. In *Proceedings COLING 2008*.

Vickrey, D. & Koller, D. (2008). Sentence simplification for semantic role labeling. In *Proceedings of the 46th Meeting of the Association for Computational Linguistics*.

Zhang, Jie, Dan Shen, Guodong Zu, Su Jian and Chew-Lim Tan (2004). Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37, 411-422.