Journal of Computing and Information Technology - CIT 15, 2007, 4, 295–303
doi:10.2498/cit.1001136

Information Extraction: The Power of Words and Pictures

Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium

The paper stresses the importance of automatically analyzing and semantically annotating creative forms of human expression, among which are textual sources. A number of challenging and emerging research directions are enumerated and illustrated with results obtained by the research group of the author.

Keywords: text analysis, cross-media and cross-lingual alignment and understanding

1. Introduction

We humans create content and the digital age has given us wonderful tools to generate masses of it. We like to explore this creativity when expressing how we perceive the world around us. Our *natural language* is the first, important example. Although our utterances are structured to a certain degree, there are a myriad of ways of how we can express content by combining an almost finite set of words and syntactic constructs. The power of understanding lies in recognizing the patterns of communication and interpreting them according to our linguistic, cultural and other background knowledge, and in making inferences, when not all content is made explicit.

In a *multimedia* environment, we have even more possibilities of expression. We add illustrative pictures to our texts, we shoot videos or add suggestive music to a sequence of movie scenes. In our society, these very creative forms of expression have an influence on our convictions, political opinions and societal relationships, that is often underestimated. We humans have no trouble with aggregating the different media and inferring messages and interpretations from them.

Although English has become a global language, people and nations use many other languages in communication, creating a need for cross-language understanding and aggregation of content, especially in business settings where products are sold over the whole world. In addition, languages change dynamically and efforts to standardize language do not seem to work as people prefer to use dialect forms that are characteristic of a certain *community* (e. g., in Web logs). Written and speech *data,* especially in an electronic context, is notorious for being incoherent and full of grammatical and spelling errors that are drafted with (e. g., spam messages) or without purpose (e. g., instant messages, postings on informal news groups). But, we humans cope with it without any apparent problems.

In professional settings, one has recognized the need for structuring information so that it can be easily retrieved and used in decision making. Many *knowledge management systems* have been set up. We see a recent trend into natural, pictorial representations of a knowledge domain. Quite *informal* representations such as topic maps become popular. Again, we humans do not have difficulties in interpreting them and extracting from them the correct patterns that are used in a decisive process.

What do all these forms of human expression have in common? They are highly *unstructured.* Unstructured does not imply that the data is structurally incoherent, but rather that its information is encoded in a way that makes it difficult for machines to immediately interpret it. When humans create content, the result is often far away from a logical and structured

representation that would be easy to process by the machine. We create content in an intuitive and natural way. Our creations are often full of ambiguities and we might express similar content in a variety of ways. But, our human brain manages to understand and to interpret the content, because it makes the right contextual disambiguation, associations and inferences.

Citizens, professionals, governments and companies need automated tools to search, mine and synthesize these unstructured data. Given their huge amounts, we need the machines to process them. Extracting information from the data is a very important first step. We define *information extraction* as follows [15, p. 226]:

Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, images, audio and video, providing additional aids to access and interpret the unstructured data by information systems.

Extraction is *adding meaning to content*. In the case of textual content, this means adding meaning to a term, phrase, passage, or to a combination of them, notwithstanding the many variant expressions of natural language that convey the same meaning. For instance, we detect that a certain name is a person name or name of an organization, or we extract from the text that company *X* acquired company *Y* on date *Z* for an amount *W*. When adding meaning to pictures, we recognize in them certain persons or objects, or classify them as outdoor scenes.

The aim of this paper is to give an overview of a number of important techniques for information extraction from unstructured data with a large focus on information extraction from a text, to illustrate them with our own current research and to reveal important points of attention for future research. The rest of this paper is organized as follows. We will first discuss current technologies for information extraction from texts, where we discuss extraction from single sentences towards extraction across sentences and documents. A very new and promising area of research is text analysis for annotating images and the recognition and mutual reinforcement of content across different media (e. g., images and texts). Then follows a section on information extraction from blogs, community texts and other natural utterances. Finally we discuss the applications of information extraction in information search, mining, synthesis and information visualization.

2. Information Extraction from Well-formed Texts

Information extraction from a text has quite a long history. The works of Roger Schank [18] and Marvin Minsky [13] in the 1970s are very important in this respect. They taught us that content in a text is composed of small elements which the author of the text has combined in order to communicate a certain message. A strong impetus for developing information extraction technology came from the *Message Understanding Conferences* (MUC), held in the 1980s and 1990s, currently succeeded by the *Automatic Content Extraction* (ACE) competition. Another solid stimulus for developing information extraction technology currently originates from the *biomedical field* where content becomes only manageable with the help of this technology. The third important factor regards the growing use of techniques of content recognition in *multimedia*.

There are a number of typical information extraction tasks that have lately been extensively researched with regard to open domain information extraction and that are becoming included in commercial applications. They include named entity recognition, noun phrase coreference resolution, entity relation recognition and timeline recognition.

Named entity recognition classifies named expressions in a text (such as person, company, location or protein names). In the example "Mary Smith works for Concentra," "Mary" is recognized as a person and "Concentra" as a company. *Named entity recognition* – and more specifically recognition of persons, organizations and locations – in news texts is fairly well developed, yielding performance in terms of F-measure* above 95% (e. g., [3]). The performance of named entity taggers on written

* F-measure here refers to the harmonic mean, a measure that combines recall and precision where recall and precision are equally weighted (also referred to as F_1-measure).

documents such as Wall Street Journal articles is thus comparable to human performance, the latter being estimated in the 94-96% F-measure range. In the biomedical domain, *named entity recognition* is a very common task because of the absolute necessity to recognize names of genes, proteins, gene products, organisms, drugs, chemical compounds, diseases, symptoms, etc. Depending on the semantic class, F-measures range up to 80% (e. g., [19]).

Another important task is *noun phrase coreferent resolution*. Two or more noun phrases are coreferent when they refer to the same situation described in the text. Many references in a text are encoded as phoric references, i. e., linguistic elements that, rather than directly encoding the meaning of an entity, refer to a direct description of the entity earlier (anaphoric) or later (cataphoric) in the text. In the example "Bill Clinton went to New York, where he was invited for a keynote speech. The former president..."., "Bill Clinton", "he" and "the former president" refer in this text to the same entity. "He" refers to an anaphoric reference. This is a quite difficult task with F-measures exceeding 70% (e. g., [5]).

Semantic role recognition regards the assignment of semantic roles to the (syntactic) constituents of a sentence [10]. The roles regard certain actions or states, their participants and their circumstances. When detecting circumstances, time expressions are the most studied (see below). Semantic role detection plays also an important role in entity relation recognition. In the example, "John Smith works for IBM", the relation "employee" between John Smith and IBM is detected. Entity relation recognition receives considerable attention in the biomedical domain. The named entity recognition is a first step for more advanced extraction tasks such as the detection of protein-protein interaction, gene regulation events, subcellular location of proteins and pathway discovery. In other words, the biological entities and their relationships convey knowledge that is embedded in large textual document bases that are electronically available. Exact numbers of performance depend on the type of relation that is extracted. With sufficient training examples we attain F-measures in the mid 80% [15].

Temporal expression detection and resolution has lately received research attention [12]. The first task is timex (i. e., temporal expression) detection and classification in text expressions to be marked, and it includes both absolute expressions (e. g., July 17, 1999, 12:00, the summer of '69) and relative expressions (e. g., yesterday, last week, the next millennium). Also noteworthy are durations (e. g., one hour, two weeks), event-anchored expressions (e. g., two days before departure), and sets of times (e. g., every week). From the recognized timexes, the time line of different events can be reconstructed. Basic temporal relations are: X before Y, Xequals Y, X meets Y, X overlaps Y, X during Y, X starts Y, X finishes Y. Recognizing a time line involves sophisticated forms of temporal reasoning. For example, from the example "On April 16, 2005 I passed my final exam. The three weeks before I studied a lot.", we could extract: *March 26, 2005-> April 15, 2005: Study* and *April 16, 2005: Exam*. Detecting temporal expressions in the text is not complicated and compares to the above information extraction tasks in terms of performance numbers. Resolving the timexes into an absolute or relative time line is much more difficult.

The above extraction tasks (with the exception of relation recognition) are rather domain-independent. But, they already allow identifying many of the details of an event (e. g., time, location). Domain-dependent extraction tasks can be defined to complement an event description (e. g., the number of victims of a terrorist attack, the symptoms of a disease of a patient). At this level, information extraction regards the extraction of information about individual events (and states), the status of participants in these events and their spatial, temporal or causal setting.

In our past research, our group also worked on the above themes. We developed a named entity recognizer, and modules for coreference resolution and detected semantic roles in texts. We are using this expertise in our current research projects where an automated semantic clarification of textual content is the aim. For instance, in the CADIAL project we automatically index legislative texts.

Nowadays, *powerful computers* are omnipresent and the advancements in the processing of natural language text allow doing things that were unthinkable a few decades ago. Especially, the *availability of reliable learning systems* makes

advanced information extraction possible. Most of the techniques use supervised learning, i. e., training a system based on annotated examples. Among the most successful classification algorithms are support vector machines, maximum entropy models and conditional random fields.

In all the above cases we train a *classifier* based on annotated examples. Each example is described by a feature vector x which describes a number of features that may refer to the information element to be classified and to its context. The goal is to train a predictive classifier based on the examples and to use this learned model in order to assign a label y to a new example. Whereas in our early research the features that we used to classify the texts were just words, in our later and current research we incorporate the features obtained through the use of natural language processing tools such as a part-of-speech tagger that detects the syntactic category of a word in a text, or a sentence parser that identifies the dependencies between sentence constituents. In the AntiPhish project, we work with many other kinds of features such as salting features, features that are not rendered on the screen by an e-mail client, but are part of the content, and which mislead filters (such as very small invisible fonts) [7].

The classifiers that we choose to work with allow dealing with incomplete data because they adhere to the *maximum entropy principle*. This principle states that, when we make inferences based on incomplete information. we should draw them from that probability distribution that has the maximum entropy permitted by the information that we do have. Examples of such classifiers are the *maximum entropy model* [2] and *conditional random fields* [11]. We also work with learning techniques capable of dealing with a large set of features that on occasion might be noisy, such as a *Support Vector Machine* [6]. A support vector machine gives us also the possibility to work with structured objects instead of feature vectors (for instance, the dependency tree or an html tree of e-mails) and to define a kernel function for computing the similarity between these objects. In other circumstances, we use context-dependent classifiers, when the assignment of one class not only depends on a certain configuration of features, but also on other classes assigned, i. e., on other feature vectors of objects in the context (e. g., conditional random fields).

Machine learning techniques that learn the extraction patterns have many advantages. It is often worthwhile that a knowledge engineer acquires symbolic knowledge that can be unambiguously defined and shared by different applications. Machine learning naturally allows considering many more contextual features than is usually the case with handcrafted rules. Moreover, language is an instrument of a society of living persons. As it is the reflection of the properties, thoughts, ideas, beliefs and realizations of that society, the extraction model should *dynamically adapt to the changing patterns* of a living language. Machine learning for information extraction has still other advantages. There is a *lesser building effort* compared to extraction systems that rely on handcrafted extraction patterns. Annotation is usually considered as being easier than knowledge engineering. Moreover, the learning techniques allow a *probabilistic assignment of the semantic labels*. Because sufficient training data or knowledge rules are usually not available in order to cover all linguistic phenomena, or the system is confronted with unsolved ambiguities of the language due to content left implicit or purposely left ambiguous by the author, there is an advantage of using learning techniques that adhere to the *maximum entropy principle* (e. g., conditional random fields) in order to cope with incomplete data.

However, there is still a lot of room for improvements. An important aspect in the management of and access to unstructured sources is their annotation or indexing with complex semantic concepts, i. e., concepts that are composed of intermediate or more simple concepts. For instance, events in the real world never exist in isolation, but rather are part of more complex events that are causally linked to each other. Humans recognize these linked events as event complexes because they stereotypically occur in a certain order. We call these stereotyped event complexes *scripts* or *scenarios*. The eventual goal of information extraction at a textual level is to recognize scenarios and to link them to abstract models that reflect complex events in the real world. Also, other "complex" semantic concepts could be applied to text, like the ones referring to issues such as "liability", "competitiveness", "medical malpractice", which, like

scenario concepts, are often themselves composed of intermediary and simple concepts. For instance, the "taking the bus" scenario is composed of a person getting on the bus at location *A*, possibly paying for the ride, the bus going from *A* to *B*, and the person getting off the bus at *B*; the concept "medical malpractice" requires a disease of a patient, a wrongly chosen treatment and a consequent suffering of the patient.

The number of *semantic concepts* by which we perceive the world around us is almost infinite and the concepts change dynamically when the content of our information sources (e. g., Web content) alter. Content creators are encouraged to manually assign semantic labels to information sources, but the economic cost is high. Hence there is a need for assisting technologies, especially if we want to assign "complex" semantic concepts, which at this point in time do not yet exist. Our group starts to perform research on this matter. In the ACILA project (Automatic detection of arguments in a legal case) we go beyond factual information extraction and recognize complete argumentation structures in legal cases and political speeches [4].

3. Cross-document and Cross-media Recognition of Content

Because our recognition techniques are not perfectly accurate, it is often useful to consider evidence from many different sources. For instance, the fact that we can attribute the title "former president of the United States" to "Bill Clinton" can be evidenced in many texts, making the attribution more certain. In addition, detecting coreferring expressions (referring to persons, locations, temporal expressions, events, etc.) and linking these expressions across documents is very useful when one wants to mine the information. This is not always a simple task due to the problems cited above, i. e., many expressions (e. g., the name "Michael Jordan") are ambiguous and refer to different persons, and the expressions in which the same content is made explicit (e. g., "the trial of the Enron case" or "Enron goes to court") are multiple. There are a number of interesting algorithms that can be used for this matter. What we do is *aligning content across documents*. For instance, we are interested in the problem of noun phrase coreferent resolution across documents, and we would like to link person names including the noun phrases by which they are coreferred (e. g., in the example above, that Bill Clinton is a "former president"). If we can start from a reasonable detection of the coreferents within one document (i. e., grouping the coreferring names with an already good probability), evidence from multiple documents helps us to correct and improve these initial assignments. One possibility is, for instance, to initially assign detecting coreferring noun phrases in a text and then making these assignments more accurate by clustering the noun phrases across documents with the expectation maximization algorithm.

The need for alignment of content is also present in a *multimedia context*. We illustrate our texts with images and we create videos. If you want to search information in a multimedia archive, it is important that you find related material. There is not much research yet on cross-media alignment. We are currently working on two projects with regard to this topic. In both CLASS and AMASS ++ projects, we annotate images based on accompanying text, align content across these media and summarize video.

The central objective of the CLASS project is to develop advanced statistical learning methods that allow images, video and associated text to be analyzed and structured automatically. Because object recognition in images is a very difficult task, since the end of the last century there is an increasing interest in using textual descriptions as a weak annotation of image content. In this situation, the visual and textual information can be considered as comparable, ranging from quite parallel content pairs provided by the text and the image to a more loose correlation between the text and the image, where the text contains, for instance, much more additional information which is not in the image, or vice versa. In most current approaches, the text that accompanies an image is seen as a bag of words, ignoring that the text's discourse structure and semantics allow for a more fine-grained identification of what content might be present in the image. In our research, we have successfully integrated the knowledge about discourse structures with the semantics in our model.

We have performed experiments with image-text pairs of Yahoo news (Figures 1 and 2) and

with pairs of the video of Buffy, the Vampire slayer. We focused on the entities, i. e., persons and objects. We assume that the more salient and the more visual the entity is in the text closely associated with the imagery, the higher are the chances that the entity is present in the accompanying image and, perhaps, the more prominent is the entity in it. Firstly, *salience* is measured by hierarchically segmenting the texts into its topics and subtopics. Here, a segmentation model was trained on the DUC (Document Understanding Conference) corpora. As classifier we used a mixture model where the interpolation weights were learned with the expectation maximization algorithm [14].

Secondly, the visualness of an entity is a measure to what extent this entity can be made visible in the image [8] (Figure 1). For computing this *visualness*, we rely on additional resources that semantically classify the word. For entities expressed by common nouns, WordNet is a resource where a limited number of seed visual entities are manually recognized. The visualness of other entities mentioned in WordNet is inversely proportional with the distances to the visual seeds. For detecting the visualness of proper names, we rely on named entity recognition. When aligning persons and objects between images and accompanying texts, we could improve the F-measure by 24.63% when incorporating a visualness score, 28.21% when incorporating a salience score, and by 32.17% when incorporating a combined visualness and salience score (by considering entities with this score $>= 0.4$), compared to a baseline approach of 31.28% F-measure that solely considers the content words of the accompanying text. The integrated visualness and salience score multiplies the salience and visualness probability (assuming independence) to obtain the probability that an entity E_i (person or object) mentioned in the accompanying text T_j is present in the image, i. e., $P(E_{i\text{-}in\text{-}image}|T_j)$. The entities are ranked by this value (Figure 2). If we assume that the number of faces in an image can be more or less correctly detected by state-of-the-art technology, a cut-off of the ranked list by this number yields 95.06% F-measure of the entities actually present in the image, compared to an F-measure of 81.65% if we use the nouns in the texts sorted by their position, which for the short texts is already a valid baseline. We have also demonstrated that the ranking obtained with our method correlates with the importance of persons and objects in the image.

Hiram Myers, of Edmond, Okla., walks across the fence, attempting to deliver what he called a 'people's indictment' of Halliburton CEO David Lesar, outside the site of the annual Halliburton shareholders meeting in Duncan, Okla., leading to his arrest, Wednesday, May 17, 2006. (AP Photo)

Visualness of fence: 0.79
Visualness of indictment: 0.0
Visualness of shareholders: 1.0
Visualness of meeting: 0.0
Visualness of arrest: 0.0

Figure 1. Example image-text pair from Yahoo! News illustrating the visualness score.

Currently, we study the *recognition* of visual *attributes* of persons and objects, and the *actions* in the texts in which the persons or objects are involved. We acquire visual attributes (e. g., visual adjectives) from large descriptive and non-descriptive corpora. We learn the visualness of a word from a corpus of texts that with a large certainty describes the images and from a normal reference corpus, while testing the hypothesis that the word and its visualness class occur independently. Chi-square, or a likelihood ratio for a binomial distribution, provides a natural way to reject the hypothesis of independence with a certain probability. We can then assume that the complement to this probability represents the visualness of the entity.

Sen. Hillary Rodham Clinton, left, and her husband, former President Bill Clinton stand at the start of a memorial service for former senator and Treasury Secretary Lloyd Bentsen in Houston, Tuesday, May 30, 2006. (AP Photo/LM Otero, Pool)

Entity	**Visualness**	**Salience**	**Combined**
Bill Clinton	1	0.975	0.975
President	1	0.975	0.975
Hillary Rodham Clinton	1	0.75	0.75
Lloyd Bentsen	1	0.3876	0.3875
Secretary	1	0.3875	0.3875
husband	1	0.25	0.25
senator	1	0.175	0.175
start	0	0.4642	0
memorial	0	0.4166	0
service	0	0.4166	0
May	0	0.3416	0
Tuesday	0	0.2916	0
Houston	0	0.1607	0

Figure 2. Example image-text pair from Yahoo! News resulting in the ranking shown in the third column based on $P(E_{i\text{-}in\text{-}image} \mid T_J)$.

In this first model we annotate the images with relevant text. The annotation is only used to build a content model of the image. If we have many image text pairs, the accuracy of the alignment can be improved by this evidence. The work of Berg et al. [1] is along these lines. These authors improve the alignment between faces in images and person names in the captions by clustering with the expectation maximization algorithm. Eventually, our vision partners in the CLASS project will classify persons in images, even in the absence of texts, based solely on the visual patterns, but our text analysis has provided them with a very useful training set.

Our approaches also allow detecting the content in the text that is not present in the image. In these cases, both contents can compliment each other, for instance, when generating a synthesis across the media sources.

4. Cross-lingual Recognition of Content

We use many different languages for communicating content. They include official languages and the myriad of *community languages* (pictorial languages, dialects, messaging languages, blogs), the latter especially being used for ventilating the voice of the people.

Confronted with standard languages, we are able to align content in parallel corpora, i. e., when one corpus is the exact translation of the other. But, parallel corpora are found in rather artificial situations, for instance, they might constitute governmental documents in the official language of a state. Much of our multilingual information is available in comparable corpora. These corpora treat similar content, but are not exact translations of each other. Research only starts to align information in comparable documents (e. g., [17]). In our AMASS++ project, we want to tackle the problem of cross-lingual alignment in comparable corpora, in addition to the cross-media alignment.

There are many unofficial languages. In blogs, we see typical examples. Because they lack any standardization, they are much more difficult to process. We have experienced this phenomenon in several projects. In the A4MC[3] project, we worked with dialect texts from the Belgian city of Hasselt [9]. Extracting information from these texts is a big problem. We cannot rely on part-of-speech taggers or sentence parsers. Standard words and dialect words are mixed with each other; words are written in many different spellings, making the paraphrasing problem even worse. In the project *Time-based text analytics*, on which we work together with the company Attentio of Brussels, we extract sentiments and opinions from blogs [4]. Here, we are especially confronted with noisy

texts (examples in Figure 3), and content is arranged in all types of formats. This situation gets even worse in the project AntiPhish, where we process spam mails [7]. Spammers become more and more inventive in hiding and obfuscating content, which misleads spam filters. The sources that we process here are very different from well-formed natural language texts.

Meanwhile, humans have created other forms of communication like hypertexts, wikis and topic maps. And who knows what else?

> pr ceu ki conaise pas le trip en vrè, pr ns, bmw c pa une voiture c un mec mè jvè pa metre son nom XD!
> (topic: BMW)
>
> 2006 MOVIE REVIEW A Good Year a flat bouquet Nothing but a French kiss-off Gladiator collaborators seem defeated by lightweight love story. By ROBERT W.
> (topic: A Good Year)

Figure 3. Example blog texts that express an opinion with regard to the topic car "BMW" (in French) and the movie "A good year" (in English) respectively found on the World Wide Web.

5. Applications

The assignment of meaning to unstructured sources and the alignment of content across different documents, media and languages is a prerequisite for effective *information retrieval*, *mining*, *summarization* and *presentation of content*. One can imagine many applications. Insurance companies are interested in having a multimedia summarizer of accident accounts. Electronic files of court cases are a conglomerate of texts, video and speech records that need to be efficiently and effectively searched and mined. Police and intelligence services are very much interested in mining repositories of images and texts among which are police reports. Many e-learning settings are of multimedia nature (e. g., in the medical domain) and students want course material presented and tailored to their needs and capabilities. Who would not want a personalized virtual museum tour or a personalized video of the evening news captured from different broadcasters? And there is the large World Wide Web stuffed with news, images, video, blogs, news groups texts and people's opinions, a barometer of society one wants to mine, monitor and search. These example applications demonstrate an absolute need for the development of the technologies discussed above.

6. Conclusions

Although there are many initiatives to standardize and structure our communicated data at the origin, in order to make them more manageable by the computer (initiatives of the semantic Web), we see that just the opposite happens. People start to invent a myriad of new forms of communication, which are often very different from rigid standard forms. It is as if the story of the tower of Babel is happening for a second time.

It is an important challenge for the machine to understand the many different ways by which humans create the content, certainly providing research topics for the years to come. To describe the content with meaningful concepts and to align it across media and languages will remain challenging tasks. The most successful techniques will be the ones that are highly adaptive to novel content formats and languages. Hence the usefulness of techniques that learn from few annotated data or that in an unsupervised way exploit patterns reoccurring in large data sets. This research story is only beginning.

7. Acknowledgements

We are very grateful to the organizations that sponsored the research projects mentioned: ACILA (Automatic Detection and Classification of Arguments in a Legal Case), K.U.Leuven Onderzoeksfonds (OT/06/03); AMASS++ (Advanced Multimedia Alignment and Structured Summarization), IWT (SBO 060051), in collaboration with: K.U.Leuven (Visics, Luc Van Gool, CCL, Frank Van Eynde), Universiteit Hasselt (EDM); AntiPhish (Anticipatory Learning for Reliable Phishing Prevention), EU FP6-027978, in collaboration with: Fraunhofer-Gesellschaft zur Fördering der angewandten Forschung (IAIS, Bonn), Germany, Symantec LIRIC Limited, Ireland, Symantec Ltd, USA, TISCALI Services, Italy, Nortel Networks, France;

CADIAL (Computer-Aided Document Indexing for Accessing Legislation), Ministerie van de Vlaamse Gemeenschap, in collaboration with: University of Zagreb (Croatia); CLASS (Cognitive-Level Annotation Using Latent Statistical Structure), EU FP6-027978, in collaboration with: K. U. Leuven (ESAT-Visics, Luc Van Gool), INRIA, Grenoble, France, University of Oxford, UK, University of Helsinki, Finland, Max-Planck Institute for Biological Cybernetics, Germany; TIME (Advanced Time-Based Text Analytics) IWOIB, in collaboration with: Attentio, Belgium.

References

[1] T. BERG ET AL. Names and faces. *Technical report*, University of California Berkeley, (2007).

[2] A. D. BERGER, S. A. DELLA PIETRA, V. J. DELLA PIETRA, A maximum entropy approach to natural language processing. *Computational Linguistics*, Vol. 22, No. 1, (1996), 39–71.

[3] D. M. BIKEL, R. SCHWARTZ, R. M. WEISCHEDEL, An algorithm that learns what's in a name. *Machine Learning*, **34** (1999), 211–231.

[4] E. BOIY, P. HENS, K. DESCHACHT, M.-F. MOENS, Automatic sentiment analysis of on-line text. *Proceedings of the 11th International Conference on Electronic Publishing, Openness in Digital Publishing: Awareness, Discovery & Access*, (2007) Vienna, Austria.

[5] C. CARDIE, K. WAGSTAFF, Noun phrase coreference as clustering. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora ACL*, (1999) pp. 82–89.

[6] N. CHRISTIANINI, J. SHAWE-TAYLOR, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK, Cambridge University Press, 2000.

[7] J. DE BEER, M.-F. MOENS, A general solution of (hidden) text salting. *Techical Report*, (2007).

[8] K. DESCHACHT, M.-F. MOENS, Text analysis for automatic image annotation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, (2007) East Stroudsburg, ACL.

[9] W. DE SMET, M.-F. MOENS, Generating a topic hierarchy from dialect texts. In *Proceedings of the 4th International Workshop on Text-based Information Retrieval (TIR-07). IEEE Computer Society.*

[10] D. GILDEA, D. JURASKY, Automatic labeling of semantic roles. *Computational Linguistics*, Vol. 28, No. 3, (2002), 245–28.

[11] J. LAFFERTY, A. MCCALLUM, F. C. N. PEREIRA, Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *Proceedings of the 18th International Conference on Machine Learning.*, (2001), pp. 282–289. Morgan Kaufmann, San Francisco, CA.

[12] I. MANI, J. PUSTEJOVSKY, R. GAIZAUSKAS, (Eds.), *The Language of Time: A Reader*. Oxford, Oxford University Press, 2005.

[13] M. MINSKY, A framework for representing knowledge. In *P. H. Winston (Ed.)*, (1975) pp. 211–277. The Psychology of Computer Vision, McGraw-Hill, New York.

[14] M.-F. MOENS, Using patterns of thematic progression for building a table of content of a text. *Journal of Natural Language Engineering*, Vol. 12, No. 3, (2006), 1–28.

[15] M.-F. MOENS, *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series 21)*. Springer, New York, 2006.

[16] M.-F. MOENS, E. BOIY, R. MOCHALES PALAU, C. REED, Automatic detection of arguments in legal texts. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law*, (2007) New York, ACM.

[17] D. S. MUNTEANU, D. MARCU, Extracting parallel sub-sentential fragments from comparable corpora. *Proceedings of ACL-2006*, pp. 81–88, (2006) Sydney, Australia.

[18] R. C. SCHANK, *Conceptual Information Processing*. North Holland, Amsterdam, 1975.

[19] J. ZHANG, D. SHEN, G. ZU, S. JIAN, C. L. TAN, Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, **37** (2004), 411–422.

Received: June, 2007
Accepted: September, 2007

Contact address:
Marie-Francine Moens
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B-3000 Heverlee, Belgium
e-mail: Marie-Francine.Moens@cs.kuleuven.be

MARIE-FRANCINE MOENS is an associate professor at the Department of Computer Science of the Katholieke Universiteit Leuven, Belgium. She holds a M. Sc. and a Ph. D. degrees in Computer Science from this university. She currently leads a research team of 10 researchers and Ph.D. students who study topics of text-based information retrieval. Her main interests are in the domain of automated content retrieval from texts, using a combination of statistical machine learning and symbolic techniques and exploiting insights from linguistic and cognitive theories.