Multi-Paradigm and Multi-Lingual Information Extraction as Support for Medical Web Labelling Authorities

Martin Labsky[1], Vojtech Svatek[2], Marek Nekvasil[3]
[1]IBM TJ Watson Research
[2,3]Dept. of Information and Knowledge Engineering
University of Economics, Prague
Czech Republic

Abstract: *Until recently, quality labelling of medical web content has been a pre-dominantly manual activity. However, the advances in automated text processing opened the way to computerised support of this activity. The core enabling technology is information extraction (IE). However, the heterogeneity of websites offering medical content imposes particular requirements on the IE techniques to be applied. In the paper we discuss these requirements and describe a multi-paradigm approach to IE addressing them. Experiments on multi-lingual data are reported. The research has been carried out within the EU MedIEQ project.*

Key words: extraction ontology, information extraction, medical web labelling.

1. Introduction

The number of web sites with medical content, on which patients as well as professionals seek advice, is steadily increasing, but their quality is rather variable. For patients and general public it is rather difficult to assess the reliability of information on the web; a solution thus is to let professional agencies, *labelling authorities*, assign quality labels to websites that fulfil important *quality criteria*. Examples of such criteria are the presence of contact information for healthcare professionals responsible for the medical content, or clear separation of publicity material from editorial material.

As medical websites typically consist of numerous pages and their structure is quite diverse, the work of labelling authorities is quite demanding. In order to collect information needed for qualified decision about accreditation, it is typically necessary to browse a large part of the website, scroll down long pages of text, and carefully examine the wording of certain paragraphs. Although some parts of this work require human common sense, there are several more routine subtasks that can be successfully accomplished by a machine.

As most of the relevant information has the form of text, *information extraction* (IE) technology seems quite at hand. Information extraction is usually understood as technology for automated discovery of (references to) various entities and relationships among them in textual data. It has been extensively tested on web data, such as product catalogues, online reviews, weather forecasts, seminar announcements and the like. The underlying techniques are quite heterogeneous: from 'deep', model-based, natural language processing, through heuristic induction of symbolic rules, to statistical techniques relying on large volumes of labelled corpora. The destination for the extracted data is typically a database (where extracted entities fill the fields in a record template) or an ontology (thus being populated with instances of concepts and relationships).

The crucial assumption of our research was a sophisticated IE tool could be capable of pre-selecting reasonably-sized bits of textual information in the websites, which would guide the labelling expert towards the assessment of a particular quality criterion, and sometimes even directly suggest the proper value for the criterion.

2. State of the Art

As said above, medical web accreditation has traditionally been addressed by medical bodies and considered as a demanding manual activity. Later on, computerised support has been added to this effort, however, it rather focused on the phase of relevant document retrieval, as in [2] and [3]. The key issue in these systems was to *distinguish* medical resources from non-medical ones, and the

supporting technique for this was the detection of medical terms, using machine-readable nomenclatures such as MeSH. *Detection* of MeSH terms was also used for website quality assessment in [4].

A somewhat different aspect of medical websites was investigated in the RankMed project [5]. There the main goal was to analyse the appearance and ergonomics aspects of medical websites, such as the visibility and validity of links, presence of language versions, site map or date of last update.

The research described in this paper obviously differs from the retrieval-oriented projects by its focus on fine-grained information extraction, which aspires to alleviate the labelling authorities from a significant part of the manual scrutiny of relevant web pages. The difference from the RankMed lays in its openness: while the RankMed quality criteria are hard-wired in proprietary software, our software architecture described below allows to almost arbitrarily extend and modify the scope of website content analysis, either by re-designing an extraction ontology or by inductively training a new extraction model from data and plugging it into the architecture.

3. Challenges of Medical Web Labelling for IE Tools

Information extraction is a heavily investigated field in the intersection of knowledge engineering and natural language processing. It has already proven very successful in two strikingly different settings: named entity recognition and wrapper-based web IE.

In *named entity recognition* (NER), the goal is to extract lexical items that refer to instances of a limited number of real-world entity types: the most frequent ones are people, organisations, locations, currency, dates and the like. The extraction here typically relies on large lists of known lexical items, so called *gazetteers*, plus context/content models inductively trained on large labelled corpora. NER tools typically perform well for free texts, especially those from general domains such as newspaper articles, as the training corpora usually belong to these domains as well. They however frequently fail in narrow domains and for semi-structured text (such as HTML pages), where the context of to-be-extracted information often does not conform to the standard structure of natural language sentences. NER tools also do not by themselves allow combining elementary lexical items into structured records corresponding to complex entities (e.g. persons with contact information). Although the underlying statistical or symbolic *inductive methods* in principle enable to build models for other than 'traditional' entities, the cost of building such models is (often prohibitively) high for real-world applications, as it first requires to manually label a sufficient amount of texts.

Wrapper-based information extraction, in turn, primarily sticks to the regularity of HTML code (or to the visual layout of this code when rendered). The semantics of a particular information item is derived from the position in which it appears in the HTML presentation structure. This leads to relatively easy extraction of structured data records; however, for a single website (built according to the same template) only. A new wrapper has to be designed, manually or with partial help of inductive procedures, for every resource.

It can easily be seen that the extraction of information relevant for medical website accreditation is a different 'species' of IE than those typically addressed by either NER tools or wrapper tools:

- There is a large number of *heterogeneous* resources with varying size, degree of structuredness and domain specificity; identical semantic types of data are 'packed' differently from one resource to another
- Off-the-shelf NER tools (equipped with models trained on generic corpora) cannot be directly used for the purpose, as information relevant for accreditation often does not fall under the 'classical' *categories* (typically, they are finer-grained: special categories of persons, organisations etc.); moreover, such NER tools only exist for a few major languages
- Manual *collection* and *annotation of training data* would be very tedious in many cases.

An adequate approach to address this challenge seems to be a *hybrid* one. Some types of information are relatively scarce and 'buried' in large portions of text (which makes their manual annotation hard) but have rather predictable content and context; they can thus be captured by *manually formed* patterns. An example from the medical domain are privacy, advertisement, sponsor policy and similar statements. Some others have highly variable content and context but can more easily be identified in the pages themselves in abundant quantities; for such entities it makes sense to *manually label* their occurrences for *inductively training* an extraction model. For example, names of medical professionals having a scientific degree belong to this category. Finally, even if the *HTML structuring* of information differs from website to website and sometimes (for semi-manually built sites) even from page to page, there are *local regularities* that are worth capturing: if several table cells in the same row/column or

several items in a list are known to contain information of the same semantic type, this type can also be deduced for the remaining items.

As the extraction relies on multiple, even heterogeneous, sources of extraction evidence, a substantial method of *confidence management* is also needed. For example, a sequence of tokens that

- does not resemble any known person name
- but appears in a table column containing some known person names
- and under a column header cell containing an indicative context token (e.g. 'members')

should be proposed as potential name, but with lower confidence than a sequence of tokens at least partially matching a relevant gazetteer (under the same circumstances).

This led us to the development of a framework and tool for information extraction that relies on combination of handcrafted patterns, trained models and 'local wrappers'. Furthermore, extraction evidence is combined in a pseudo-probabilistic manner. Finally but importantly, the extraction patterns, rather than being simply stored in a database, are appended to a high-level data model equipped with logical constraints. We conceive this model, which plays the key role in connecting our IE approach with the semantic web paradigm, as an *extraction ontology* (cf. [6]).

4. IET and *Ex*: Multi-Paradigm Information Extraction Framework

4.1. IET as part of AQUA

Within the AQUA architecture for supporting medical website quality evaluation [7], the information extraction subsystem, nicknamed as IET (Information Extraction Toolkit), is one of the central ones. The task of IET is to extract standalone named entities and also instances composed of related attributes. The 'real work' of extracting information items is carried out by the IE engines included in IET. Currently there are two such engines in IET: a conventional *statistical tool* based on conditional random fields (CRF) and our own *Ex* engine, which has *extraction ontologies* as its central principle. The remaining components of IET either have infrastructural character such as the Task Manager, Data Model Manager and Document I/O subsystem, or have special purposes like the Evaluator and the Annotation tool (serving for manual labelling of data for training and testing).

Although IET can operate as a completely stand-alone IE application, its functionality is currently used within the framework of AQUA, where it interacts (through a standardized API) with other components; see the diagram of architecture in Fig. 1. The documents subjected to extraction have already been processed by WCC – the Web Content Collection toolkit; this processing includes finding relevant websites on the Internet, spidering for individual pages and automated assignment of page category. The page category is then exploited within IET for finer-grained classification of extracted results. The results of extraction and automatically annotated documents are then sent forth to the label database and to the "Monitor-Update-Alert" (MUA) component.

4.2. The *Ex* information extraction tool

The approach to web information extraction implemented in *Ex* was originally inspired by that developed by Embley and colleagues [6]. The most important distinctive features of *Ex* compared to its pre-cursors are:

- The possibility to provide extraction evidence with *probability estimates* plus other quantitative info such as value distributions, allowing to calculate the likelihood for every attribute and instance candidate using pseudo-probabilistic inference [8]. The most likely sequence (through the processed document) of instance candidates and standalone attribute values is finally extracted.
- The effort to combine hand-crafted extraction ontologies with other sources of information: *HTML formatting* and/or *training data*. HTML formatting is exploited via *formatting pattern induction*. Training data can be exploited via incorporating external *inductive learning* tools.

Extraction ontologies in *Ex* are designed so as to extract occurrences of attributes (such as 'person name' or 'organisation name'), i.e. standalone named entities or values, and occurrences of whole *instances of classes* (such as 'contact information'), as groups of attributes that 'belong together'. A *specialisation hierarchy* can be defined at the level of attributes and classes, such that various pieces of evidence can indicate e.g. specialisation of 'contact information' to 'medical responsible contact information', 'administrative person contact information' or 'web designer contact information'.

The whole *extraction ontology definition language* is quite rich; it allows specifying e.g. nested regular patterns for attribute content and context (in terms of characters and words, HTML tags as well as labels provided by external tools), scriptable axioms that need to hold for the extracted values, co-reference resolution rules, formatting constraints or cardinality ranges for attributes within classes. Finally, scripted rules may be applied to the set of extracted attribute values and instances in order to post-process and finalize their values based on arbitrary information such as document category. A more comprehensive description of this language can be found in [9], and an exhaustive reference is in the tutorial available from [10].

Ex ontologies are particularly suited to extract *classes* and their *attributes* from web pages that exhibit some *formatting structure*, but can also be applied to *free text*. Relation extraction is only limited to determining the membership of attribute values in instances of extractable classes. On this basis, *Ex* has been successfully tested in the domains of *online product offers*, *weather forecasts* or *seminar announcements* (an overview is e.g. in [12]). The medical accreditation task fits well into this category, as the individual labelling criteria are relatively independent of each other and its goal is to automatically provide the labelling expert with rich information relevant to a given criterion so as to allow him/her to form his/her opinion without having to delve into the plethora of pages of the given website.

4.3. Other components and aspects of extraction

As mentioned above, *Ex* is able to exploit labels assigned to text by other IE tools. Currently we use a *pre-processing IE engine* whose aim is to detect named entities corresponding to contact information using a CRF-trained model from a human annotated corpus of documents. In general, *conditional random fields* (CRFs) [11] are a probabilistic framework for labelling and segmenting sequential data using a conditional probability distribution over label sequences given a sequence of words. We have chosen CRFs because they outperform both “Maximum Entropy Markov models” and “Hidden Markov Models” on a number of real-world tasks in many fields, including bioinformatics, computational linguistics and speech recognition. In the current IE engine, the CRF++ software (http://crfpp.sourceforge.net/) is used.

An important aspect of the current research is *multi-linguality*. IET is being tested with six languages: English, Spanish, Czech, German, Greek and Finnish. Localization to new languages is relatively easy as is the inclusion of new sets of *labelling criteria*. Simple guidelines for extending the extraction for further criteria and languages (including e.g. the modification of extraction ontology, collection and labelling of data for the new criterion/language, and training a classifier) and more details on IET are described in [12].

5. Experimental Results on a Multi-Lingual Corpus

5.1. Overview of experiments

In this section we present selected results of extracting *contact information* attribute values in six different languages using the extraction ontology engine and the CRF engine; complete results can be found in [13]. For English, German and Czech we also present initial results for instance extraction.

Apart from contact information, we also implemented the task of extracting selected *free-text criteria*. We however do not evaluate this task here as we lack sufficient amount of test data for this domain.

It is also important to state that quantitative evaluation is only a partial means of evaluating the feasibility of the approach. Namely, the extraction task should be viewed in the context of the labelling authority day-to-day work, and the added value should be measured by the improvement of this work. This *user-centric evaluation* was carried out for the whole AQUA architecture in parallel to purely technological evaluation, and its outcomes are described in [14].

5.2. Contact information extraction task definition and experimental setting

Contact information is one of the critical and versatile features in medical (or other) website accreditation: its absence can be *decisive* for the quality label assignment, and it appears in various species in *various labelling criteria*, e.g. as ‘medical content responsible contact information’, ‘administrative person contact information’ or ‘web designer contact information’.

Tab. 1 describes the six document sets in different languages used to conduct the experiments. Each set consists of HTML pages collected by the WCC and manually classified as positive contact pages. All documents in these sets were annotated manually to acquire labelled gold standard data.

The following attributes constituting contact information are distinguished:

- degree-title (medical title, e.g. MD),
- job-title (e.g. manager, head of department)
- name (person name, e.g. J. Smith, John Smith, Smith),
- street (street name incl. number, e.g. 22 Oak Lane),
- city (city name, e.g. New York City, NYC),
- region (geographical units other than city or country, e.g. counties or boroughs),
- zip (zip code, e.g. 162 00, 94105-2099),
- country (e.g. Denmark),
- phone (phone numbers including extensions, e.g. +420 463 928 281),
- email (e.g. smith@uep.cz, smith at uep dot cz),
- organization (full organization name, e.g. Medical Center at the York Hospital),
- department (name of a unit, e.g. Department of Preventive Medicine).

5.3. Development of extraction ontologies

In order to build the extraction ontologies for *Ex* IE engine, 30 documents were randomly selected from each language to become available to the extraction ontology developer for viewing as "training data". The remaining documents from each set were used as test data for evaluation. Six contact extraction ontologies were developed (one per language with shared common parts). Each ontology contained about 100 textual patterns for the context and content of attributes and also for the single extracted 'contact' class, attribute length, data type constraints and several axioms. The effort spent on developing and tuning the ontologies was about roughly 2-3 person-weeks for the initial, English ontology, and 1-2 person weeks for its customization to Spanish, Czech, German, Greek and Finnish.

5.4. Combining the extraction ontology and CRF engines

The IET uses a pipeline to combine several IE engines in one extraction task. For contact extraction we first perform extraction using the CRF engine. The produced named entities are then utilized by the extraction ontology engine, which uses them as one piece of evidence indicating the presence of a named entity.

In order to evaluate the combined system, we created the following test setup. We could not use a standard cross-validation due to the fact that the extraction ontologies were created by a human expert based on seeing a fixed sample of training documents. Therefore, for each language, we created a training/test split with 80% of documents used for training and 20% used for testing. The 80% training sets include all (30) documents viewed by the expert in order to create the extraction ontologies. The CRF engine was trained using the 80% training set for each language.

The results of combining both IE engines are encouraging as they always indicate an improvement in the overall precision, recall and F-measure for each language. Also – apart for a few exceptions – the F-measures for each attribute improve as well.

In Tab. 2 we list the results for English; we omit detailed results for the remaining five languages due to space limitations (see [13] for full listing). For the same reason we only list the F-measure (and not precision and recall), and only the *strict testing results* (requiring exact correspondence of the tokens annotated). The last column in Table 2 shows the number of gold-standard annotations present in the test set. Regarding the empty results for the 'job' attribute: for English, Spanish and Finnish, this attribute is not included in the gold standard annotations as it was not part of the extraction criteria at the time of acquiring the collections. The 'region' attribute was left out from the automatic IE task as it showed high inter-annotator disagreement and it is only listed for completeness.

For the English results we can see a consistent improvement in F-measure on average compared to extraction ontologies alone. The CRF engine actually integrates easily with the extraction ontology due to its high precision: the ontology regards the CRF positive classification as a highly sufficient but not necessary piece of evidence. In particular for attributes that proved hard to model using the extraction ontology, like organisation names, department names or street names (where there was little intuition to base manual modelling on), the improvement compared to using the extraction ontology alone is relatively high. For a single attribute, name of organisation, CRF slightly outperformed the extraction ontology, but the hybrid approach was a clear winner.

Furthermore, in Tab. 3 we include meta-level results for all six languages: the counts of different F-measure comparison results between the approaches, computed on the 12 attributes. Table 4 then lists the same results but aggregated over languages, for different attributes. The first five rows correspond to *monotonic combination*, which does not lead to decrease of the F-measure in the hybrid approach. For some languages and attributes we could however also observe non-monotonic behaviour.

Results of this meta-analysis can be used to determine the optimal approach for each attribute and to tune parameters that affect how the CRF engine's decisions are used within the extraction ontology.

The complete (basic, not meta-level) results of attribute extraction are in [13]. They include a more thorough discussion and:

- *Precision* and *recall* figures from which the F-measure was computed
- An alternative gold standard matching approach. While in the *strict mode* of evaluation (referred to in the table above), only exact matches were considered to be successfully extracted, in the *loose mode*, partial credit was also given to incomplete or overflown matches. For example, extracting 'John Newman' where 'John Newman Jr.' was supposed to be extracted would count as a 66% match (based on overlapping parts).
- Individual tests of the extraction ontologies and CRF with *different settings*, such as the classical 10-fold cross-validation mode for CRF, which could not be used when making the direct comparison with extraction ontologies.1
- The above items not only for English but also for the other *five languages*: Spanish, German, Czech, Finnish and Greek.

5.5. Extraction of contact information instances

In order to increase the usefulness of extraction results for the end-user (labelling authority), extraction of isolated entities is followed with the assembly of instances; in this case, instance of class *Contact info*. We show experimental results for English, German and Czech in Table 5. Each contact instance could consist of multiple attribute values (named entities), such as name of person, his/her title, e-mail address, phone, job position, department, city, country and the like. As in the previous experiments, we distinguish the evaluation results for the extraction ontology only and for extraction ontology combined with the CRF classifier (used at the level of named entities).

The quality of instance extraction cannot be computed simply from the proportion of correctly and incorrectly extracted instances, as the degree of containment of correct named entities also matters. Therefore we computed the so-called *villain precision*, *villain recall*, and *villain score* as its combination (analogous to F-measure), see [15]. In this approach, all named entities in an instance are assumed to be pairwise connected by virtual links (e.g. 2 links are needed to connect 3 attribute values together), and the precision/recall is computed in terms of such links (considering links in the gold standard vs. links in the extracted instance). We can see that roughly two thirds of links get are identified correctly. With the CRF engine included, the improved accuracy of attribute value extraction translates to an average improvement of 1% point over the three analysed languages.

5.6. Free-text criteria extraction task

The list of labelling criteria includes several other criteria that are also addressed by the IET. They are defined as:

- Purpose/Mission of the resource
- Statements declaring sources of funding (sponsors, grants, advertisers)
- Explanation on how personal data (visitor coordinates, e-mails etc.) is handled.

These are different in nature from the contact criteria since the underlying resources are mostly free-text, meaning the units to be extracted are not named entities but typically whole sentences or blocks of text such as paragraphs or item lists. Another complicating factor is that no resources were available within the project to assemble and manually label gold-standard documents that could be used for training or testing. Apart from this, it seems that collecting significant amount of training data for each of the three free-text criteria for the needed languages may not be possible, as e.g. for Czech there may not be enough distinct examples of sponsor policies found in online medical pages.

[1] Note that we cannot apply cross-validation on the human designer of the extraction ontology, as s/he is not as 'oblivious' as a machine. An alternative could be to include multiple ontology designers in the evaluation.

With no labelled documents at hand, we decided to only use the extraction ontology engine also to extract the needed free-text criteria. An extraction ontology that performs free-text extraction relies on identifying keywords or key-phrases that appear in significant amounts in consecutive sentences or within a block of text. For example, based on observing several sponsor policies in English, we selected several tens of words or short phrases that might often appear in sponsor policies described on the web. Regular extraction patterns of the extraction ontology language are then used to extract sequences of sentences or text blocks that contain enough of these keywords and key phrases. As for contact information, free-text extraction ontology was developed first for English (in approximately 1.5 weeks) followed by its ports to Spanish and German.

6. Discussion

The experimental results suggest that the approach is viable. While achieving satisfactory results for the 'generic' *named entity extraction* task (e.g. person name or organisation name) and instance extraction (contact-info), its advantage is its applicability even in cases where little or no training data are available, due to the possibility of combining multiple sources of extraction evidence (manual, trained and regular formatting evidence). The approach based on extraction ontologies also allows for good *maintainability* and *extensibility* with further extractable items and languages.

The *hybrid approach*, in which extraction ontologies are combined with trained models, is in a majority of cases superior to both extraction ontologies alone and to trained models alone, in terms of F-measure and often both precision and recall.

An important added value (not only for the domain of medical web accreditation but also for information extraction research in general) also lies in the transition from *generic entities* to *domain-specific* ones, using both *local information* from text and *context information* external to the actual IE task such as document category.

7. Conclusions

The work presented attempts to give insights into the technological side of supporting medical web labelling authorities with fine-grained information required for convenient evaluation of individual quality criteria. The proposed hybrid approach to web information extraction combines manually formulated evidence in extraction ontologies with evidence gained from annotated corpora using an inductive tool.

Acknowledgments

The work presented in this paper was supported by the EC-funded project MedIEQ (www.medieq.org), under the DG-SANCO Programme "Public Health". Training of the CRF models for named entity recognition which were used within extraction ontologies was done by Aris Theodorakos and his colleagues from NCSR[2]. Numerous individuals from within this project consortium were also helpful in preparing the data collections and reference annotations.

References

[1] MARVIN: how does it work? Report by Health On the Net Foundation. Geneva, Switzerland, 2002. Online at <http://www.hon.ch/Project/Marvinpspecificities.html>.

[2] Griffiths K.M., Tang T.T., Hawking D., Christensen H.: Automated assessment of the quality of depression websites. J Med Internet Res. 2005 Dec 30;7(5):e59.

[3] Wang Y., Liu Z.: Automatic detecting indicators for quality of health information on the Web. Int J. Med Inform. 2006 May 31.

[4] Kim W., Aronson A.R., Wilbur W. J.: Automatic MeSH term assignment and quality assessment. Proc AMIA Symp. 2001; 319-23.

[5] Adla T., Kasal P., Hladikova M., Janda A., Naidr J., Feberova J., Kubu P., Potuckova R.: Use of the automated quality evaluation system for the comparison of health care web pages. In: MedNet 2004.

[6] Embley D. W., Tao C., Liddle D. W.: Automatically extracting ontologically specified data from HTML tables of unknown structure. In: Proc. ER '02, pp. 322--337, London, UK, 2002. Springer-Verlag.

[2] National Center of Scientific Research "Demokritos"

[7] Karkaletsis V., Karampiperis P., Stamatakis K., Labsky M., Ruzicka M., Svatek V., Mayer M. A., Leis A., Villarroel D.: Automating Accreditation of Medical Web Content. In: 5th Prestigious Applications of Intelligent Systems Conference (PAIS 2008), Greece. Incl. in Proc. ECAI'08, IOS Press, 2008.

[8] Labsky M., Svatek V.: Combining Multiple Sources of Evidence in Web Information Extraction. In: 17th International Symposium on Methodologies for Intelligent Systems (ISMIS'08), Toronto, May 20-23, 2008. Springer LNCS 4994, pp. 471-476.

[9] Labsky M., Svatek V., Nekvasil M., Rak D.: The Ex Project: Web Information Extraction using Extraction Ontologies. In: Proc. PriCKL'07, ECML/PKDD Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery. Warsaw, Poland, October 2007.

[10] Ex information extraction system. Online at <http://eso.vse.cz/~labsky/ex/>.

[11] Lafferty J., McCallum A., Pereira F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning, pp. 282-289.

[12] Labsky M., Svatek V., Nekvasil M.: Information Extraction Based on Extraction Ontologies: Design, Deployment and Evaluation. In: 1st International and KI-08 Workshop on Ontology-based Information Extraction Systems (OBIES 2008), online <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-400/>.

[13] MedIEQ Deliverable D9.2: Final version of information extraction toolkit. Online at <http://www.medieq.org/results>.

[14] MedIEQ Deliverable D16.1: Evaluation of the First Prototype. Online at <http://www.medieq.org/results>.

[15] Moens M. F.: Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer, 2006. ISBN 1-4020-4987-0.

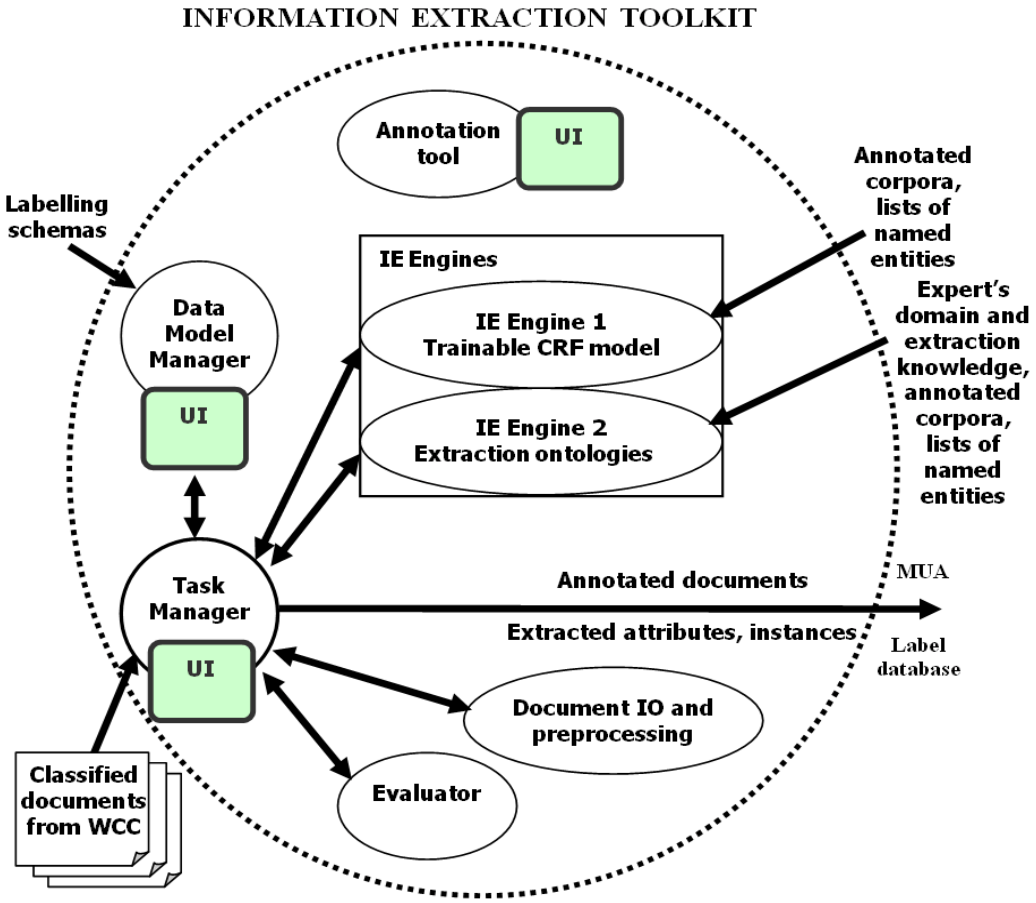

Fig. 1. The final IET architecture and data flow

Tab. 1. Document and annotation counts for positive contact pages in six languages.

Language	Documents	Named entities
English	116	7000
Spanish	206	5000
Czech	99	11000
German	93	4950
Greek	106	4300
Finnish	98	5692

Tab. 2. F-measures for the separate and combined results of the Ex extraction ontology engine and the CRF IE engine using a 20% test split of the English contact dataset.

Attribute	*Ex*	CRF	*Ex*+CRF	Difference	Gold count
City	49.9	26.3	**50.1**	+0.3	127
Country	72.6	51.1	**74.2**	+1.6	101
Degree	76.3	58.7	**77.3**	+0.9	276
Department	57.9	54.4	**62.6**	+4.7	71
Email	100	100	**100**	0	55
Job	-	-	**-**	-	0
Name	71.5	52.3	**73.4**	+2.0	828
Organization	45.5	46.8	**50.6**	+5.1	342
Phone	**67.7**	0	66.3	-1.3	107
Region	14.4	2.0	**14.6**	+0.2	97
Street	58.0	39.1	**66.7**	+8.6	35
Zip	89.5	65.5	**89.5**	0	36
Overall	*64.3*	*48.9*	***66.3***	*+1.9*	*2075*

Tab. 3. Counts of different F-measure comparison results, for all languages

Result	EN	ES	DE	CZ	GR	FI	Total	Perc.
Hybrid > EO > CRF	7	3	2	3	2	2	**19**	**26.4%**
Hybrid > CRF > EO	1		2	2	1	2	**8**	**11.1%**
Hybrid = EO > CRF	1	1	3	1	5	1	**12**	**16.7%**
Hybrid = CRF > EO						2	**2**	**2.8%**
All 100%	1						**1**	**1.4%**
Total monotonic	*10*	*4*	*7*	*6*	*8*	*7*	***42***	***58.3%***
EO > Hybrid > CRF	1	3	2	1	4		**11**	**15.3%**
CRF > Hybrid > EO		2	1	2		2	**7**	**9.7%**
CRF > Hybrid = EO			1				**1**	**1.4%**
EO > CRF > Hybrid				1			**1**	**1.4%**
Total non-monotonic	*1*	*5*	*4*	*4*	*4*	*2*	***20***	***27.8%***
Not addressed by EO		1					**1**	**1.4%**
Not annotated in data[3]	1	2	1	2		3	**9**	**12.5%**

[3] In one case (degree attribute in Finnish data) there were only two entities annotated; this was treated as 'not annotated'.

Tab. 4. Counts of different F-measure comparison results, for each attribute

Result	City	Country	Degree	Dept.	Email	Job
Hybrid > EO > CRF	2	1	2	2		1
Hybrid > CRF > EO	2					1
Hybrid = EO > CRF		1		1	5	2
Hybrid = CRF > EO	1			1		
All 100%					1	
Total monotonic	*5*	*2*	*2*	*4*	*6*	*4*
EO > Hybrid > CRF	1	1	2			
CRF > Hybrid > EO				2		
CRF > Hybrid = EO		1				
EO > CRF > Hybrid						
Total non-monotonic	*1*	*2*	*2*	*2*	*0*	*0*
Not addressed by EO						
Not annotated in data		2	2			2

Result	Name	Organis.	Phone	Region	Street	Zip
Hybrid > EO > CRF	2		3	1	3	2
Hybrid > CRF > EO		2	2		1	
Hybrid = EO > CRF				1		2
Hybrid = CRF > EO						
All 100%						
Total monotonic	*2*	*2*	*5*	*2*	*4*	*4*
EO > Hybrid > CRF	3	1	1		1	1
CRF > Hybrid > EO		3			1	1
CRF > Hybrid = EO						
EO > CRF > Hybrid	1					
Total non-monotonic	*4*	*4*	*1*	*0*	*2*	*2*
Not addressed by EO				1		
Not annotated in data[4]				3		

Tab. 5. Preliminary results of contact instance extraction for three languages

Language	Gold count of instances	Villain score (*Ex* ontology)	Villain score (onto+CRF)	Improvement
English	963	63.9%	65.2%	+1.4%
German	769	58.8%	61.8%	+3.0%
Czech	1681	66.5%	65.1%	-1.5%

[4] In two cases (degree attribute in Finnish data and region attribute in Czech data) there were only two and five entities annotated, respectively; this was treated as 'not annotated'.