Service-oriented Information Extraction

Katja Seidler
SAP Research Dresden
SAP AG, Germany
katja.seidler@sap.com
Advisor: Prof. Dr. Dr. Alexander Schill

ABSTRACT
The growth of unstructured data being available in the Internet, private and enterprise file shares demands flexible and generic solutions to extract the included information. Current information extraction solutions focus on domain specific approaches manually built together and adapted for a preassigned task. Moreover they suffer from uncertainty and inaccurate results, especially when the domain slightly changes. In this paper, we propose a new service-oriented information extraction approach to overcome the present disadvantages. We outline the major topics involved in the development of such a system and point to research questions we want to address in the doctoral work.

1. INTRODUCTION
Today, a vast amount of information is unstructured and can not be understood by machines. Information Extraction (IE) is essential to make use of this knowledge in web pages, office documents, tweets and many other data sources. Extracting information from natural language documents with the help of suitable techniques is the major task of IE. Extraction tasks are for example the classification of documents (e.g., document D is a product report) and the recognition of entities (e.g., person, city, organization) and their relationships (e.g., person X lives in city Y) as well as the recognition of sentiments (e.g., product Z is well received).

Current IE systems are mostly manually developed for a certain domain in an expensive manner. Initial approaches have been developed to modularize this process. Thereby, the IE process is modeled as a pipeline that describes how things have to be extracted. However, a lot of manual work is necessary to combine the single modules and create domain specific applications. We intend to establish a new service-oriented approach to this problem. Our goal is to use IE functionalities as services, make them automatically detectable and combine them to solve IE tasks. In contrast to existing techniques such as [7], the extraction shall be started with a description of ‘what’ has to be extracted.

The envisioned system will help to make IE immediately available and easier to use. It will make use of existing IE services like OpenCalais and AlchemyAPI (see Table 1 for details) and combine them in a generic manner. By developing special algorithms the quality of the extracted information and the adaptability of the system can be increased. Due to the universality of the system it can be used in all scenarios requiring extraction steps.

The rest of the paper is structured as follows. Section 2 presents the research problem and the main challenges being addressed in the doctoral thesis. The related work is discussed in Section 3 and distinguished by our planned contributions. Section 4 gives a deep insight in the proposed approach and the topic of the thesis. The planned experimental evaluation is presented in Section 5. Finally, Section 6 gives an outlook on the next steps.

2. RESEARCH PROBLEM
IE has never been an easy task, especially not when dealing with unstructured heterogeneous text data from different sources with miscellaneous content and of varying domains. It is a manageable task to create an extractor for a specific document with a well defined content, even though its an expensive and time consuming process. But once the document, the included content or even the domain slightly changes, the extraction system will not be as reliable as before. The doctoral thesis focuses on a flexible, easy to use, reliable and generic service-oriented solution for IE. In the following part, the challenges of such an IE solution are introduced, followed by the main contributions of the doctoral thesis.

2.1 Challenges
We must face a number of challenges to create a flexible, generic, but also reliable and accurate IE solution. The first issue is the question of how to make existing IE solutions reusable to save time and costs for the expensive process of generating fitting IE systems. Related to this problem is the issue of describing IE modules precisely, regardless of them being services or software assets. As IE solutions can, in most cases, not be reused as is, an effective combination of modules is required. Besides the questions of automatically selecting and combining the IE modules, another challenge arises: how can a system aggregate the results of different, potentially overlapping IE solutions? Is it possible to raise the quality and reliability of the results by combining partly redundant systems? How can an iterative learning process...
even enforce these effects? The challenges being worked out in the above mentioned issues lead to the following goals:

Goal 1 (Reusability) Develop techniques to make IE modules reusable. This includes the definition of global interfaces, a syntactic and semantic description of the modules, as well as providing the modules in a (public) registry and ensuring that they can be searched.

Goal 2 (Combination) Develop algorithms and patterns for the goal-driven selection and combination of IE modules. The heuristics being modeled should consider the IE task, the entered data, the user preferences, but as well the cost and quality metrics being related to the IE modules.

Goal 3 (Aggregation) Develop methods to aggregate the results of different IE modules. This includes a global uncertainty management based on the results of single modules, the derivation of information, as well as the evolution of quality metrics for the IE modules and an iterative refinement of the selection, combination and aggregation process.

2.2 Thesis Contributions
The thesis proposes a new service-oriented IE approach. In contrast to most of the existing IE approaches we do not aim to create new IE systems from scratch. Rather we want to reuse existing solutions and effectively combine them task- and quality-driven. In detail, the thesis will make the following contributions to the state of the art:

- Definition of a description language for IE services and tasks, that has the ability to precisely identify the services, its interfaces, its functionalities and accordingly the semantic meaning and the requested functionality of the tasks.
- Development of algorithms for the task- and quality-driven selection and combination of semantically described services for a given IE task.
- Development of heuristics for the aggregation of single IE results from various services to a global result.
- Development of a model for uncertainty and reliability in service-oriented IE solutions.

3. RELATED WORK
This section discusses the related work of the doctoral thesis. The first part (Section 3.1) focuses on IE systems and specific solutions. Service-oriented systems and their application for IE systems are studied in the second part (Section 3.2).

3.1 Information Extraction
There have been great efforts in IE over the last decades. Especially concrete IE approaches in the field of rule-based and machine learning techniques [15] have been proposed. Tutorials [2, 5] and surveys [15] recap the effort being made. But they also point out the following requirements not having been resolved yet: (i) scalability, (ii) accuracy and reliability as well as (iii) the usability and flexibility of IE solutions.

Adaptive IE approaches [16] address the wish to offer solutions which can readjust to new domains on the basis of a training process with external knowledge. As they need sufficient training sets, a lot of manual work is still necessary. In the following paragraphs some heavily studied aspects of IE are presented.

Uncertainty IE will for sure never be a 100% accurate job. However there is a need for very accurate and reliable results especially in the business and medical domain. Therefore in addition to the refinement of the IE techniques, it is necessary to calculate reliable confidence estimates for the extracted information. Several approaches studied the problematic aspect of uncertainty in the IE process and developed probabilistic models for the reliability of extracted information in rule-based [13] and probabilistic systems [3]. The reliability of such estimates in supervised learning approaches was analyzed by [14]. The authors revealed the big error rates and inaccuracy of such information and tried to adjust the systematic failure of various IE approaches with a calibration function. Nevertheless, the problem that such confidence estimates are domain-specific and constrained by many factors remains. Therefore, these approaches could serve as a valuable input for our envisioned service-oriented IE approach, but must be extended to integrate and take into account the domain specific factors, as well as the dissimilarity of individual extraction results for the same information. One opportunity to increase the quality of confidence estimates is the determination of the influence of redundant extraction results. Initial approaches examined the effects of redundancy [6], but to the best of our knowledge there is no system that takes into account the results of different (partly overlapping) IE systems.

Modularization An essential contribution to the modularization and reusability of IE processes is supplied by the prominent Unstructured Information Management Architecture (UIMA) [7]. Modules for input and output, as well as those for analyzing and processing are manually assembled in an extraction pipeline in order to be executed during runtime. A global infrastructure allows the reusability of IE modules, but still demands technical knowledge and a manual and time consuming process to build an extraction pipeline. Instead of a definition how extraction has to be done, we focus on a system that only needs the information what to extract and automatically handles the how. Further, it is even conceivable to use the service-oriented system in a way, where even the information what to extract is derived from the user context and the entered documents.

3.2 Service-Oriented Systems
Service-oriented Systems and the world of web services are heavily explored topics with a lot of standards having been implemented (see [18] for an overview). In the next paragraphs, we will mainly investigate the research on the field of web service description and its usage in IE.

Service Description Different web service description languages exist to describe services regarding their functionalities, the used data types, the protocols and the provided interfaces. The W3C’s standard Web Services Description Language (WSDL) [1] was established for the syntactic description of services. More recent approaches [10] such as the On-
Service-Oriented Information Extraction

The first service-oriented IE approach is discussed by Habegger et al. [8]. In this connection, they break down the IE process into single operations, like parsing, filtering and extracting, to create flexible IE web services. IE services are then allocated to one of the specified operations and can be described with their proprietary Web Extraction Task Description Language (WetDL) [9]. The language is used to specify the service name, the possible successive operations and specific characteristics of the operation types. In contrast to the WetDL that does not offer a semantic description and is not generic, we will develop a semantic description language for IE web services and offer a generic solution. In [17] the first system for a service-oriented integration of text mining services is presented. The approach models text mining and knowledge extraction processes in the biomedical domain as independent distributed services. The extraction task is handled by base services with a standardized interface and a central access point, the meta service. The integration of individual results is done with the help of an aggregation service that offers routines for the visualization and consolidation of conflicting results. The presented system is a great entrance point for the thesis, but mainly lacks in the following: (i) services are not identified and connected automatically based on their quality and functionalities, (ii) the aggregation is not done on the basis of a global uncertainty management and does not consider redundancy, (iii) no learning process and iterative refinement is included in the system and (iv) there are no algorithms for the matching of extraction types.

4. PROPOSED APPROACH

The aim of the thesis is to create a service-oriented IE system that benefits from existing IE solutions and effectively combines them to guarantee flexibility in the IE process as well as quality of the extraction results. Meeting the challenges discussed in Section 2, the following hypotheses shall be proven:

1. Reusability: IE systems can be reused by making them accessible in a registry as IE services with a definite syntactic and semantic description.
2. Discoverability: Suitable services, if available, can be found with the help of the registry for a precise semantic description of a IE task and additional user preferences.
3. Dynamic Combination: By the dynamic combination of the retrieved services, IE tasks can automatically be resolved in an adequate manner.
4. Quality Intensification: The quality of the IE results can be increased through applicable heuristics, algorithms and learning processes for the selection and combination of the services.

The following paragraphs will give a deeper insight into the planned system. Therefore we first provide a system overview in Section 4.1 and present two possible application scenarios in Section 4.2. A number of the considered extraction services are shortly introduced in Section 4.3. The two major planned contributions - making IE solution reusable and improving the quality of IE results - are discussed in more detail in Section 4.4 and 4.5.

4.1 System Overview

Figure 1 shows the proposed service-oriented IE system. The extraction services are offered by a set of service providers (1) through a detailed description of the service functionalities and the provided interfaces, protocols and data types.
Table 1: Overview of existing IE services

<table>
<thead>
<tr>
<th>name</th>
<th>domain</th>
<th>type</th>
<th>feature</th>
<th>language</th>
<th>limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIIAGMT[^1]</td>
<td>biomedical</td>
<td>x</td>
<td>disambiguation, Linked Data, quotations</td>
<td>en, fr, es, de, it, pt, ru, sv</td>
<td>30,000 calls/day</td>
</tr>
<tr>
<td>AlchemyAPI[^2]</td>
<td>generic</td>
<td>xx</td>
<td>upload own specification corpus</td>
<td>en</td>
<td>max. 75 concepts/call</td>
</tr>
<tr>
<td>BeliefNetworks[^3]</td>
<td>generic</td>
<td>x</td>
<td>Linked Data</td>
<td>en</td>
<td>50,000 transactions/day</td>
</tr>
<tr>
<td>Evri[^4]</td>
<td>generic</td>
<td>x</td>
<td>precise specification through query</td>
<td>en</td>
<td>5000 calls/day, 1 call/s</td>
</tr>
<tr>
<td>OpenAmplify[^5]</td>
<td>business, finance, generic</td>
<td>x</td>
<td>max. 1MB/call</td>
<td>en</td>
<td>non-commercial</td>
</tr>
<tr>
<td>OpenCalais[^6]</td>
<td>business, finance, generic</td>
<td>x</td>
<td>max. 1MB/call</td>
<td>en</td>
<td>non-commercial</td>
</tr>
<tr>
<td>PIE[^7]</td>
<td>biomedical</td>
<td>x</td>
<td>max. 1MB/call</td>
<td>en</td>
<td>non-commercial</td>
</tr>
<tr>
<td>uClassify[^8]</td>
<td>generic</td>
<td>x</td>
<td>training</td>
<td>en</td>
<td>100-3000 signs</td>
</tr>
<tr>
<td>Yahoo Term Extraction[^9]</td>
<td>biomedical</td>
<td>x</td>
<td>training</td>
<td>en</td>
<td>5000 calls/day, only</td>
</tr>
</tbody>
</table>


4.2 Application Scenario

In order to motivate such a system presented above we will now provide two possible application scenarios.

**Market Analysis** The amount of up to date data on the Internet offers a great opportunity to analyze the current market through news mining on the Internet. Monitoring and analyzing the news in feeds, blogs, forums, newspaper articles, company websites, etc. involves several text mining tasks. For example sentiment analysis can be used to extract any subjective information from the news and on the basis of this draw conclusions for business strategies. Relationship extraction can further be used to react on extracted facts and actions. As the requirements involved in such news mining tools are rapidly changing a dynamic and adaptable system is necessary that combines different text mining and IE tools. Further the user of such news mining tools (e.g. managers) demand easy and directly to use systems and make high demands on the quality of the results.

**Life Sciences** Scientist face the problem to overlook a vast amount of potentially relevant scientific documents and publications. Specialized text mining applications help them to handle the data and extract the necessary information. Particularly for the extraction of genetic background of diseases IE has established. Powerful algorithms and tools are already existing and often offered by web services. Each of them is either adjusted to enable high precision or recall and is sometimes even specialized for certain species. As there is no one-suits-all text mining tool [17], the combination of several of such tools respectively web services promises great gains. Again the users are not skilled and even do not have the time to manually create specific analysis pipelines. Due to the accountability of the extracted genetic information and the conclusions drawn thereof the IE quality is as well an important factor.

For both scenarios presented here our envisioned system is suitable. It is on the one hand easy to use and automatically integrates different available solutions and therefore fits to the potential users. On the other hand it targets the quality improvement of IE and and is domain independent.

4.3 IE Services

The envisioned IE system benefits from existing services. At the moment, the number of (publicly) available IE services is limited, but we believe that once the infrastructure develops, the number of services as well as the attractiveness to provide them will raise. One of the first extraction services provided to the public for free was the OpenCalais service from Reuters. The service allows the automatic annotation of arbitrary text content with meta data. It identifies entities, events and facts in the provided content and additionally assigns topics and social tags to it. As a starting point for our work we analyzed existing IE services, as well as APIs and systems providing IE functionalities. Table 1 gives an overview on publicly available extraction services. Due to space limitations, only a fraction of the existing services can be presented here. We classified the services along the offered extraction types and the domain they were originally designed for. Additionally, some limitations for the service application are mentioned. The extraction types we identified are: (1) Named Entity Recognition, (2) Extraction of Entity Relationships and Interactions, (3) Categorization, (4) Concept Assignment, (5) Keyword Extraction and (6) Sentiment Analysis. Initial experiments with some of the IE services are described in Section 5.1.

4.4 Improvement of Usability

We pointed out that the existing approaches lack usability of IE solutions. We will provide techniques to improve the reusability and flexibility. First of all, as already discussed, a precise description of the extraction services is needed. We already studied the information necessary to support this task. Besides describing the syntactic characteristics about the input and output data format, we identified the following extraction specific information: (i) type of extraction (e.g., classification), (ii) supported languages, (iii) if available the domain the solution was defined for, (iv) the supported information types (e.g., the classification types), (v) if available, the used extraction methodology, (vi) support of
special features like Linked Data support or normalization of the extracted types. In addition a description about service charges and limitations, as well as service provider characteristics and privacy issues (e.g., storing the input data and extracted information at the service provider) can complete the description. For the definition of the IE task two cases are feasible: (i) describing the task with the help of the description methodologies used for the services and (ii) deriving it from training data or the provided documents, the user context and the domain.

Besides the description of the tasks and services, the task- and quality-driven selection and combination is another challenge to be solved. Therefore, we take into consideration a cost, a utility and a quality function. The aim is to find the optimum between them - low costs, with maximum task coverage and a high quality. The cost function is defined by the service charges (e.g., concrete prices, maximum transaction budget for a service), the expected service operating times, the required local resources and the local operating times. The utility of a service selection and combination is measured with the task coverage. The quality function includes the domain-specific precision and recall of individual extraction services, the estimated global precision and recall and the service quality (e.g., availability, reliability) based on those considerations, algorithms for the selection and combination process will be defined.

4.5 Quality Intensification
We believe that the quality of IE, respectively recall and precision of results, can be improved through the combination of extraction services. In the following we formally describe the idea behind this. Let us assume that we have a document $d$ with an included knowledge $K$ and $n$ extraction services $s_i (i = 1, 2, ..., n)$ that can extract a knowledge $K_i$ from the document $d$. Each knowledge $K_i$ is composed by a true knowledge set $K_i^t$, with $K_i^t \subseteq K$, and the set $K_i^f$ ($K_i^f \not\subseteq K$) of knowledge being extracted false positive. The described setting is illustrated in Figure 2.

![Figure 2: Combination of services](image)

We know that extraction services are in the most cases not 100% accurate and even do not find all the included knowledge. That means that on the one hand, we have to handle the case that a service $s_i$ only extracts a subset of $K$ and misses some knowledge $K_i^m = K \setminus K_i^t$. On the other hand, we have to assume that not all knowledge being extracted is true and we have to decide which extracted information belongs to the true knowledge $K_i^t$. In both cases the combination of several IE services, not extracting exactly the same knowledge, could help to minimize the problem. The overall missed knowledge $K^m = K \setminus \bigcup_{i=1}^{n} K_i^t$ of combined services will be smaller as far as there is not one super service $s_j$ with $K_j^t \supseteq K_i^t$ for each $i = 1, 2, ..., n$. Through the combination of services and the assumption of a certain independence between the services, the decision whether an extracted information belongs to the set $K$ can be reduced on the basis of redundancy and conflicting results.

Considering the above mentioned vision, we are planning to develop heuristics for the decision what belongs to the right knowledge and create a global uncertainty management to reflect the recall and precision of combined results. Moreover, we also have to face the fact that we need to map the results of different used terminologies onto each other and handle cases were the extracted knowledge of different services matches only partially.

5. Evaluation
In this section we present the planned evaluation of our approach. We will start with the evaluation settings and then describe the used methodology. In Section 5.1 we will give some initial results of experiments with IE services. For the evaluation, existing extraction services will be used. As we will face the problem of non-standardized interfaces, we will even have to encapsulate some of the services in an additional web service. We want to use datasets from different domains to show that the generic approach is effective. To achieve a comparison with existing approaches we will mostly rely on existing benchmarks like the one from the message understanding conferences (MUC) and the Text Retrieval Conferences (TREC).

The evaluation shall mainly answer the following three questions: (i) Can IE services automatically be found for an IE task? (ii) Can IE services be combined reasonable with respect to the context and adapt to modified user preferences? (iii) Can the quality of IE be improved by the combination and aggregation of different services and their results? The first question will be answered by testing a set of task specifications against a set of service descriptions. The task specifications will, as described in Section 4.4, either be provided directly through a description or indirectly by tagged training data and/or document characteristics. The combination of services is evaluated with the help of test scenarios that explicitly require a combination of services to be solved. A metric for the selection and combination of services will be the percentage enclosure of the functionalities and the consistency with the user preferences. To clarify the third evaluation question the results of individual services are compared with the results of the combined and aggregated service results. Finally, the gain in flexibility and saving of time will be evaluated in comparison with classic IE approaches.

5.1 Service Evaluation
We gained first results for the evaluation of the services to proof the idea behind the envisioned service-oriented IE approach. Goal of the experiments was to show that there is capability to enlarge the extracted knowledge and improve the quality of the extraction through the combination of several IE services. Additionally we wanted to prove whether the service results and the overlaps between the services depend on the domains they are used in. Therefore we ran three publicly available extraction services for named entity recognition (OpenCalais, AlchemyAPI, Evri) on sample
sets of the two corpora Enron-Mails and Reuters-News (we randomly selected 10 documents of both corpora). We manually checked the results of the three services for true and false positives, as there is, to the best of our knowledge, no benchmark for a great variety of entities.

Figure 3 depicts the results of the overlaps between the true positives of the extracted entities for the three services clustered on the individual services. The clustering was done by comparing each of the three services to the others and counting the cases, where the positions of a tagged entity matched or overlapped to one or more in the other services. The entity name and the semantic meaning of the entity was not taken into consideration for this experiment. Figure 3a) shows that 48% of all the extracted entities (of the true positives) for the Reuters corpus were found by OpenCalais. This 48% can be splitted into 6% having only been found by OpenCalais (O), 14% by Evri (E) and OpenCalais (O-E), 3% by AlchemyApi (A) and OpenCalais (O-A) and 25% by all services (O-A-E). We can see that there are differences between the two corpora. Evri performed very well for the Reuters corpus and extracted 88% of the knowledge the three services can extract. For the Enron corpus OpenCalais was the service that extracted the most information (63%).

Besides the overlaps in the result sets we analyzed the true positive rates $p$, the rates $r_m$ of exact matches compared to overlapping or included matches and the rates $r_t$, where the entity names were exactly the same for a match. The results, shown in Table 2, are again clustered on the used services and the corpora they were used on. Additionally we provided the precision numbers $p$ split on the special cases that only some services extracted the same information. It seems that there are interesting correlations between the precision and the coincidence of service results. Further experiments should go on with the evaluation of the precision on an entity level to yield better results. The values $r_m$ and $r_t$ also make clear that it is very important to provide mappings between the extracted entities and develop algorithms for decisions on not exact matches. From the experiments we can conclude that:

- the IE services show significant performance differences on diverse domains,
- the recall can be increased through the combination of several IE services,
- it often happens that the the service results are not exactly overlapping and that there are differences between the entity names that have to be resolved when combining different IE services
- and finally that the precision of IE results correlates with the coincidence of service results and it seems that with qualified algorithms for the combination and selection of the right results, the precision can be increased as well.

6. CONCLUSION AND FUTURE WORK

This paper is the starting point for our work on service-oriented IE. We proved the need for a flexible, generic, easy to use and reliable IE system and pointed out the challenges being associated therewith. We then described our first ideas and distinguished our approach from the related work. A description of the envisioned system, the presentation and classification of existing extraction services, as well as the illustration of our ideas for the selection, combination and aggregation gave a first insight into the work having been done so far. Finally, we presented the planned evaluation of our work and proved the vision behind the thesis with the help of some experiments with existing extraction services. Our next steps will focus on more enhanced experiments with such IE services. Especially we want to evaluate more functionalities and provide an extensive survey of the available services. We are further on the way of defining an initial IE description language and develop an global uncertainty model for service-oriented IE.

7. REFERENCES


<table>
<thead>
<tr>
<th>$p$</th>
<th>$r_m$</th>
<th>$r_t$</th>
<th>O</th>
<th>A</th>
<th>E</th>
<th>O-A</th>
<th>O-E</th>
<th>O-A-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>O(Reuters)</td>
<td>0.95</td>
<td>0.92</td>
<td>0.66</td>
<td>0.79</td>
<td>-</td>
<td>-</td>
<td>0.86</td>
<td>0.96</td>
</tr>
<tr>
<td>O(Enron)</td>
<td>0.86</td>
<td>0.71</td>
<td>0.53</td>
<td>0.85</td>
<td>-</td>
<td>-</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>A(Reuters)</td>
<td>0.83</td>
<td>0.80</td>
<td>0.56</td>
<td>-</td>
<td>0.13</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>A(Enron)</td>
<td>0.93</td>
<td>0.90</td>
<td>0.56</td>
<td>-</td>
<td>0.85</td>
<td>-</td>
<td>0.99</td>
<td>-</td>
</tr>
<tr>
<td>E(Reuters)</td>
<td>0.96</td>
<td>0.88</td>
<td>0.61</td>
<td>-</td>
<td>-</td>
<td>0.96</td>
<td>-</td>
<td>0.96</td>
</tr>
<tr>
<td>E(Enron)</td>
<td>0.95</td>
<td>0.92</td>
<td>0.59</td>
<td>-</td>
<td>-</td>
<td>0.87</td>
<td>-</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 2: Detailed Comparison of IE services

---

As some services extracted areas with two or more entities included into each other, the numbers show slight differences for the different clusters.


