

Fusing automatically extracted semantic annotations

Andriy Nikolov

Knowledge Media Institute, The Open University, Milton Keynes, UK

1 Research problem

One of the necessary preconditions of the Semantic Web initiative is the availability of semantic data. The Web already contains large amounts of information intended for human users. This information is mainly stored as hypertext, which must be semantically annotated to make it accessible for software agents. The amount of information on the Web makes it impossible to solve the annotation task manually. So the usage of automatic information extraction algorithms is essential. These algorithms use various NLP and machine learning techniques to extract information from text (Ciravegna 2003; Cimiano 2005). The information extracted from different sources must then be integrated in a knowledge base, so that it can be queried in a uniform way. This integration process is called *knowledge fusion*. Semantic annotations extracted automatically will inevitably contain defective aspects, which can cause problems during integration. These defective aspects include the following (based on (Appriou, Ayoun et al. 2001)):

1. *Ambiguity*. In general, information which can be interpreted in several distinct ways is considered ambiguous. It applies to the case when it is impossible to decide what real-world item the information refers to. For instance, when dealing with geographical information a document describing a country with the name “Korea” does not allow a software agent to judge automatically, whether it refers to North or South Korea.
2. *Uncertainty*. Uncertainty refers to the case when it is not possible to say definitely whether a particular Boolean statement is true or false. For instance, information can be biased depending on the source it comes from. Also there is a possibility of incorrect extraction by the extraction algorithm. For instance, the NationByNation.com website often contains values (like “Russia’s unemployment rate is equal to 1.5%”), which are different from other sources. This information should have lower reliability.
3. *Imprecision*. Sometimes the content of the statement can be imprecise itself. For example, it is possible that a statement contains a rounded number instead of the precise one (e.g., “\$1.2 million” and “\$1212000”).
4. *Incompleteness*. In most cases an information source does not contain full information about the real-world item it describes. Some properties may be missing from the description. For instance, an entity such as a country has many properties (economic, geographic, etc.). It is unlikely that all source documents will explicitly mention all of them.
5. *Vagueness*. Sometimes a predicate of a statement can be represented by a vague term, for example, “high”, “young”, “fast” etc.
6. *Inconsistency*. Inconsistency occurs when there exists a set of mutually contradicting statements. For instance, the unemployment rate of Russia is given by the CIA World Fact Book and the NationByNation web site as 7.6% and 1.5% respectively.

These defects in information have three main origins. First, existing information extraction algorithms cannot ensure 100% extraction correctness, which leads to uncertainty and incompleteness. Second, multiple sources can contradict each other, which leads to inconsistency and ambiguity and uncertainty. Third, the information itself can be imprecise, incomplete or vague. Thus, we can say that the problems of imprecision, incompleteness and vagueness are inherent for the fusion input data. The aim of the fusion step is to solve these problems using multiple sources. In contrast, the problems of ambiguity and inconsistency (and uncertainty, which is produced by them) are caused by using multiple sources and are part of fusion process itself. Therefore we can attempt to overcome them by improving the fusion algorithm.

The goal of the research is to propose a framework, which will deal with the problem of fusing automatically extracted semantic data.

2 Current approaches

We can distinguish two main research communities working on issues related to knowledge fusion on the Semantic Web. The first one can be referred to as the “knowledge fusion community”, which addresses generic problems of information integration. The second can be referred to as the “Semantic Web community” and tries to handle knowledge fusion specifically in the context of the Semantic Web. Approaches within the knowledge fusion community are generally based on well-established formalisms like first-order predicate logic to represent information and probability theory to reason about uncertainty. Approaches within the Semantic Web community are more data-driven and often use machine learning techniques.

A general overview of knowledge fusion problems is given in (Appriou, Ayoun et al. 2001). It can be seen as a summary of research approaches from the knowledge fusion community. In one of the approaches (Gregoire 2006) inconsistency resolution is based on the usage of semaphores – additional propositions, which describe the fact that only one of the conflicting statements is true. This approach is strictly formal and allows conflicts to be resolved while remaining within the limits of Boolean logic. However, the flexibility of such an approach in representing different degrees of confidence seems limited.

The study described in (Hunter and Liu 2006) proposes to overcome this limitation and resolve inconsistencies using such formalisms as probability theory and possibility theory. This work focuses on merging structured news reports in an XML format and provides formal ways of representing inconsistent information and fusing it, even if the inconsistency is measured using different formalisms (e.g., one piece of information uses probability theory and another one possibility theory).

Another study from the same research group (Hunter 2006) suggests that the importance of inconsistencies should be evaluated. First, it proposes to use four-valued Belnap logic, which allows assigning one of four values: “true”, “false”, “unknown” or “both”. Then, it introduces two metrics to measure the degree of inconsistency within the knowledge base: concordance and coherence. And, finally, to evaluate the relative significance of different incoherent subsets of a knowledge base the authors propose to assign different weights to different subsets of knowledge bases using domain knowledge. All these metrics can then be used to decide whether to ignore inconsistencies, resolve them or reject them based on a pre-defined threshold.

As we can see, the methods from the knowledge fusion community mostly focus on resolving conflicts rather than identifying them (i.e. inconsistency problem rather than ambiguity). These approaches allow the conflicts to be resolved in a straightforward way but they rely on the availability of meta-data (e.g., distribution of weights between attributes, probability assignment etc.). It is not always clear where to get this meta-data.

In the Semantic Web community the knowledge fusion problem consists of two problems, which so far have been primarily addressed separately. One is the problem of *ontology integration*, which corresponds to the generic knowledge fusion task and deals with meta-data. The second is the *instance fusion problem* (factual data fusion), which deals with ontological instances. Currently my research is focusing on the instance fusion problem. The ontology integration problem is a relatively well established topic within the Semantic Web (Rodriguez and Egenhofer 2003; Ehrig, Haase et al. 2004; Ehrig and Staab 2004; Euzenat, Loup et al. 2004). In contrast, the instance fusion problem became important only recently with the adoption of automatic information extraction algorithms in the Semantic Web domain (Ehrig, Haase et al. 2004; Guha and Garg 2004).

An approach, which deals with a problem relevant to the instance fusion task using machine learning techniques, is the one by (Guha and Garg 2004). It handles one particular case of the instance ambiguity problem and tries to determine whether two instances representing people with the same names in fact refer to the same person or to different people. The algorithm described in the paper makes its decision based on the distance between instances’ properties. The distance calculation process is straightforward. All property values are treated as strings. The distance between the values of the same property is assigned equal to 0 if the values are equal and 1 otherwise. The distance between instances is calculated as the weighted average of distances between properties where weights of properties are assigned using domain knowledge. On the basis of this calculated distance the algorithm makes a decision about resolving the potential ambiguity.

In general, methods from the Semantic Web community mostly address the problem of ambiguity rather than resolve inconsistencies.

3 Proposed approach

As discussed in the previous section, currently the approaches to knowledge fusion problem can be divided into two groups: formal “top-down” methods from the generic knowledge fusion community and quantitative “bottom-up” techniques from the applied Semantic Web community. Both approaches have their limitations. They focus on different kinds of issues and another difference between them lies in the amount of meta-data required to deal with defects in information. Therefore it makes sense to develop a knowledge fusion framework, which will combine both types of approaches so that a knowledge fusion system can make a decision depending on the type of problem and the amount of domain information it possesses. An outline of such a framework is shown in Fig. 1 (the case of multiple ontologies is not considered). In order to construct such a framework it is necessary to investigate several research problems, which include representing meta-data, employing qualitative methods and constructing the fusion framework.

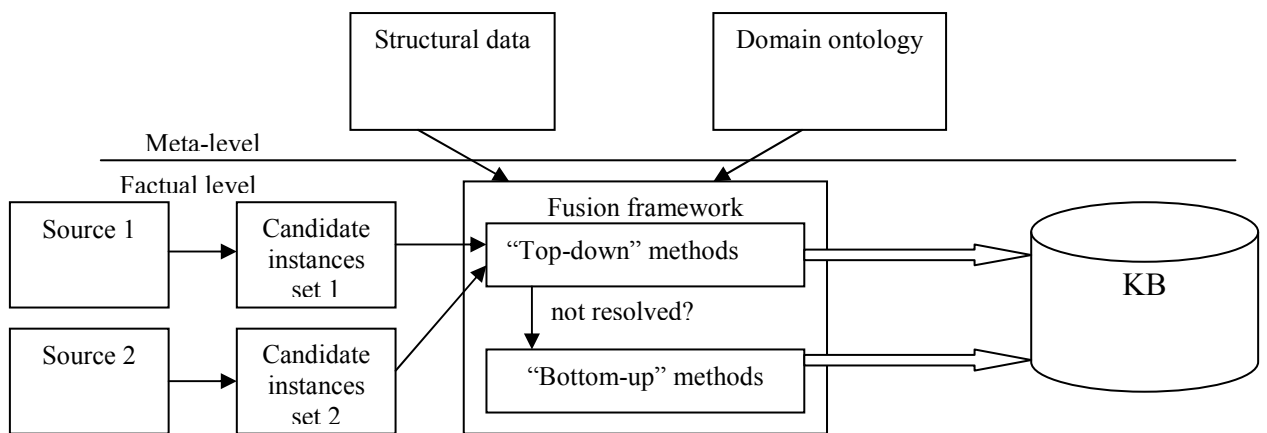


Figure 1 – Instance fusion framework

3.1 Representing meta-data

The first problem is the issue of representing the meta-data needed to deal with information defects. In the Semantic Web context domain knowledge is represented by ontologies. When tackling the knowledge fusion problem, however, it becomes necessary to describe the sources of factual data. This information, strictly speaking, does not belong to the domain description and usually is not present in knowledge bases by default. It means that we have two types of meta-data:

1. *Structural data*. When dealing with information coming from different sources the problem of the sources’ reliability arises. Therefore the knowledge base must contain explicit information about the source of each particular piece of data. This information should include at least the degree of reliability for each source and the date to which it refers.
2. *Domain ontology*. Different aspects of domain data can also become useful for solving the fusion task. These include, for example, the following:
 - explicit restrictions imposed on properties (e.g., a country’s GDP cannot be less than 1000 USD);
 - relative importance of different properties (e.g., a country’s capital name is more important for its identification than a list of trade partners).

In order to build the knowledge fusion framework the first step is to formulate the requirements, which meta-data must satisfy. The next step is to propose optimal ways to represent and store this meta-data.

3.2 Quantitative methods

The second issue deals with quantitative methods. As mentioned before, currently proposed methods for instance fusion use simple algorithm based on the calculation of similarity between instances from different sources. However, the calculation of their similarity is based on the assumption that all properties are nominal. This can be a potential disadvantage when dealing with

attributes of different types, like numeric, free text or lists. Using more flexible distance measures can improve the algorithm's performance. These potential improvements must be investigated with regard to their applicability, overall performance and robustness. These issues are discussed in more detail in (Nikolov, Uren et al. 2006).

3.3 Fusion framework

The third task is to propose a framework, which combines all techniques together to perform knowledge fusion. This task includes finding solutions to the following problems:

1. *Representing uncertainty.* Different reliability of sources, inaccuracy of extraction algorithms and quantitative fusion techniques – all these factors introduce uncertainty in the knowledge base. In order to reason about this uncertainty there must be a way to represent it explicitly in the knowledge base. Different formalisms exist, which deal with uncertainty: probability theory, possibility theory, fuzzy logic and Dempster-Shafer theory of evidence. The problems concern the choice of an appropriate formalism and how to represent all kinds of uncertainty in a uniform way.
2. *Organisation of the decision-making process.* When combining techniques of different types it is important to organise their collaboration. In our case it means choosing when fusion can be performed on the basis of meta-data and when it is necessary to rely on quantitative algorithms. Fig.1 shows a proposed framework to deal with the instance fusion problem assuming that there is a single domain ontology. The proposed way is to try to resolve the conflicts using formal “top-down” methods and, in case of failure, to resort to the quantitative algorithms. If those are also unable to solve the problem then all conflicting variants should be stored and returned in response to the user's query. The final decision thus is left to the user.

4 Distinction from other approaches

As was discussed in the “Related work” section currently there are two main kinds of approaches to the knowledge fusion problem, which we called “top-down” and “bottom-up”. The potential liability of the “top-down” approaches is their high requirements about availability of meta-data. Also the approaches found in the literature were not designed to be applied to Semantic Web: they do not consider the issues of data representation and language expressiveness assuming that data is already represented in a common format, which allows expressing all necessary data. One of the aims of our approach is to adapt these techniques to the Semantic Web domain, which means solving the issues of representing additional structural data and adopting standard ontological languages for describing this data (e.g., Fuzzy OWL (Stoilos, Stamou et al. 2005) for uncertainty representation or PML (da Silva, McGuinness et al. 2004) for provenance information).

On the other hand “bottom-up” approaches are not sufficiently accurate and robust. So the usage of meta-data should be helpful for improving their performance. Thus, our hypothesis is that a combination of formal approaches with quantitative techniques can help in cases where available meta-data is not sufficient.

5 Expected contributions

The project aims at investigating issues and making contributions in the following areas:

1. Specifying the requirements and possible solutions for the representation of meta-data needed to perform knowledge fusion (e.g., structural data, importance of attributes).
2. Investigating the robustness of quantitative machine learning methods and their applicability together with formal approaches.
3. Proposing a coherent knowledge fusion framework, which utilises emerging standards for the representation of uncertainty and dynamic knowledge, able to perform the fusion of factual data automatically extracted from different sources.

6 Evaluation

This research is a part of the X-Media project financed by EU. Results from the research will be used in industrial applications in specific domains, which means that at the evaluation stage it may be possible to use the data provided by industrial participants of the project. This makes it possible to evaluate the results based on:

1. Quantitative measures such as precision and recall when comparing to the gold standard data.
2. Qualitative feedback from the potential users.
3. Task-based evaluation.

7 Current state

This PhD project started on the 1 October 2005. So far the research has focused on two aspects: literature survey and pilot study. The aims of the pilot study are to investigate the applicability of quantitative instance fusion techniques (see (Nikolov, Uren et al. 2006) for details). Currently the experiments are still in process but some preliminary results are already available. When fusing geographical data from the CIA World Fact Book and Travel Document Systems web site (www.traveldocs.com) the algorithm with its current settings was able to identify correctly 70% of conflicting instances. The results so far do not guarantee necessary robustness when using this quantitative algorithm standalone. Further experiments will allow us to make necessary conclusions about the degree of its applicability.

8 Acknowledgements

This work is supported by the Open University (Open University Competitive Studentship Grant 881) and X-Media project (EC Grant IST-FF6-26978). I would like to thank my supervisors: Dr. Victoria Uren, Prof. Enrico Motta and Prof. Anne de Roeck for their help and guidance.

References

- Appriou, A., A. Ayoun, et al. (2001). "Fusion: General concepts and characteristics." International Journal of Intelligent Systems **16**(10): 1107-1134.
- Cimiano, P. (2005). Ontology Learning and Population. Dagstuhl Seminar "Machine Learning for the Semantic Web".
- Ciravegna, F. (2003). (LP)2: Rule Induction for Information Extraction Using Linguistic Constraints. Sheffield.
- da Silva, P. P., D. L. McGuinness, et al. (2004). A proof markup language for semantic web services. Stanford, Stanford University.
- Ehrig, M., P. Haase, et al. (2004). Similarity for Ontologies - a Comprehensive Framework. Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM 2004.
- Ehrig, M. and S. Staab (2004). QOM - Quick Ontology Mapping. 3rd International Semantic Web Conference, Hiroshima, Japan.
- Euzenat, J., D. Loup, et al. (2004). Ontology alignment with OLA. 3rd ISWC2004 workshop on Evaluation of Ontology-based tools (EON), Hiroshima, Japan.
- Gregoire, E. (2006). "An unbiased approach to iterated fusion by weakening." Information Fusion Logic-based Approaches to Information Fusion **7**(1): 35-40.
- Guha, R. and A. Garg (2004). Disambiguating People in Search. 13th World Wide Web Conference, New York, USA.
- Hunter, A. (2006). "How to act on inconsistent news: Ignore, resolve, or reject." Data & Knowledge Engineering **57**(3): 221-239.
- Hunter, A. and W. Liu (2006). "Fusion rules for merging uncertain information." Information Fusion Logic-based Approaches to Information Fusion **7**(1): 97-134.
- Nikolov, A., V. Uren, et al. (2006). Measuring similarity for fusing extracted knowledge. 9th Annual CLUK Research Colloquium, Milton Keynes, UK.
- Rodriguez, A. and M. Egenhofer (2003). "Determining Semantic Similarity Among Entity Classes from Different Ontologies." IEEE Transactions on Knowledge and Data Engineering **15**(2): 442-456.
- Stoilos, G., G. Stamou, et al. (2005). Fuzzy OWL: Uncertainty and the Semantic Web. International Workshop of OWL: Experiences and Directions, Galway, Ireland.