Department of
Information Engineering
and Computer Science DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

LARGE SCALE SEMANTIC MATCHING: AGROVOC VS CABI

Aliaksandr Autayeu

May 2011

Technical Report # DISI-11-465

Equals (modulo formatting) to the camera-ready version which will be published in the IJCAI 2011 workshop called LHD-11, “Discovery of Meaning On the Go in Large and Heterogeneous Data”, homepage: http://dream.inf.ed.ac.uk/events/lhd-11/.

Large Scale Semantic Matching: Agrovoc vs CABI

Aliaksandr Autayeu
University of Trento
Italy
autayeu@disi.unitn.it

May 22, 2011

Abstract

Achieving semantic interoperability is a difficult problem with a lot of challenges yet to address. Some of them include matching large-scale data sets, tackling the problem of missing background knowledge, evaluating large scale results, tuning the matching process and doing all of the above in a realistic setting with resource and time constraints. In this paper we report the results of a large-scale matching experiment performed on domain-specific resources: two agricultural thesauri. We share the experience concerning the above mentioned aspects of semantic matching, discuss the results, draw conclusions and outline perspective directions of future work.

1 Introduction

Semantic heterogeneity is a well-known problem and arises in many fields and applications. Many solutions have been proposed and many aspects of the problem are covered in papers (see surveys [3, 2, 15]) and books [4]. However, the problem is far from being considered solved. In fact, a definitive list of challenges yet to address in the field have been proposed [16].

This paper considers an instance of semantic heterogeneity problem: matching two thesauri. The experiment described in this paper relates to several of the problems highlighted in [16]. We test the feasibility of conducting a large scale mapping experiment. Large scale experiments were conducted before in [9] and [10], but this paper pushes the bar one order of magnitude higher: to the data sets with tens of thousand nodes. We test the importance of background knowledge with larger coverage for a domain-specific problem. Similarly to [13] we focus on agricultural domain, but on different aspects of the matching problem. This experiment can also serve as use case and as starting point for a creation of a new large-scale golden standard.

This paper describes an experiment in matching two thesauri: Agrovoc and CABI. This paper briefly describes the technique used for matching, the matched thesauri, the set up of the experiment, the experiment results and their evaluation. Finally, we make conclusions and propose future work.

2 Matching technique

In this experiment we use a semantic matching algorithm S-Match [11] empowered with a minimal semantic matching algorithm [7]. The semantic matching algorithm implemented in the S-Match framework[1] [6] consists of four steps. First, input sources, which consist of labels in natural language, are enriched with logical formulas using concepts drawn from a linguistic resource. Second, the formulas are contextualized to reflect the position of the concept in the initial data. During these two steps we use natural language processing techniques tuned to short text [19]. As a result we have transformed our input into lightweight ontologies [12]. Third, all atomic concepts identified in the source and target thesaurus, are matched using background knowledge and other techniques, like string matching. Fourth, complex concepts from source and target thesaurus are matched using satisfiability solver and axioms collected on the third step.

The minimal semantic matching algorithm MinSMatch [7] exploits the fact that given the two source trees and the mapping, the mapping may contain some elements that may be derived from the others. Such elements are called redundant and can be excluded from the mapping, reducing it, often significantly. MinSMatch allows one to compute directly such a minimal mapping, saving computational efforts.

As a source of background knowledge for the first and the third steps we have used WordNet [5]. In an attempt to alleviate the problem of lack of background knowledge [8] we also use an extended version of WordNet, made available by the Stanford WordNet project[2]. WordNet provides a good coverage of the general part of the language, its slowly changing core. In turn, the extended version of WordNet contains about 4 times more concepts than the original WordNet 2.1. For example, we extracted 78 551 (WordNet: 19 075) multiwords and 1 271 588 (WordNet: 755 306) hypernym relations. The extended version is generated automatically and is 84% accurate [18].

3 Thesauri

Agrovoc thesaurus and CABI thesaurus are two thesauri that contain terms from agricultural domain. They both cover the same domain of agriculture, which makes the matching task specific to agricultural domain. However, the knowledge base we use is generic and provides limited coverage of the agricultural domain. This factor alone is a serious challenge.

[1] `http://semanticmatching.org/`
[2] `http://ai.stanford.edu/~rion/swn/`

3.1 Agrovoc thesaurus

For our experiments we have used two versions of Agrovoc thesaurus. For the first two runs we have used hierarchical representation of the thesaurus from 10 August 2007[3] and for the second two runs — the database version of the thesaurus from 03 November 2009[4], which we converted into a compatible representation.

The version from 10 August 2007 was preprocessed and some terms[5] with a special mark-up were removed. The preprocessed version contains 35036 terms.

The version from 03 November 2009 is significantly improved in content and in structure as compared to the 2007 version. It was loaded directly from the database and converted into a hierarchical representation. Again, same terms with a special mark-up were removed. This resulted in 40 574 loaded terms. After several checks[6] on the thesauri structure and following terms removal, 32 884 terms remained. Finally, after taking into account multiple broader terms (BT) issue and decision to leave all BT links, including those leading to multiple inheritance, we have obtained a tree with 47 805 leaf-terms.

All the following figures describe the version of Agrovoc from 03 November 2009. Out of 32 884 terms 11 720 are single words, 21 161 are multiwords, that is, terms having more than one word. Most of them use space as a separator, although there are few terms using ",", "-" and "/" as a separator. Table 1 provides details on comparison of thesauri.

The final tree of 47 805 leaf-terms is a hierarchy containing a maximum of 14 levels. 76 BT relations were found to be redundant[7]. 1 207 terms have more than one BT relation. During this conversion we decided to leave these relations, effectively multiplying terms. Leaving a term in two places in a hierarchy increases the chances of it to be matched.

3.2 CABI thesaurus

For our experiments we used an "unversioned"[8] version of CABI thesaurus, which we converted into a compatible representation. The available version is not complete. As we found out during the conversion, there are many terms, referred to, but not present. We suppose that the chemical and taxonomical terms are missing.

After the same checks we did for Agrovoc, 18 200 terms were loaded. 57 BT relations were found to be redundant. 2 546 terms have more than one BT relation. During this conversion we decided to leave these relations, effectively multiplying terms. Leaving a term in two places in a hierarchy increases the chances of it being matched. The final tree of 29 172 leaf-terms is a hierarchy containing a maximum of 7 levels.

Out of 18 200 terms 6 842 are single words, 11 358 are multiwords. Most of them use space as a separator, although there are few terms using ",", "-" and "/" as a separator. Table 1 provides detailed thesauri comparison.

[3]file ag_hierarchy_20070810_EN_UTF8.txt

[4]file agrovocnew.zip/agrovocnew 20091103 1627.sql

[5]like "Siluroidei (141XXXXXXX)"

[6]symmetric use of USE/UF, hierarchy redundancy, multiple broader terms

[7]if "A nt C" and "A nt B nt C" then relation "A nt C" is considered redundant

[8]file CABTHESNontaxNonchem.txt dated from 01 November 2009

Characteristic	CABI	Agrovoc
Tree leaves	29 172	47 805
Terms count	18 200	32 884
Single words	6 842	11 720
Multiwords	11 358	21 161
Hierarchy depth	7	14
multiple BT	2 546	1 207
redundant BT	57	76

Table 1: Thesauri comparison

```
cachexia
TNR: 18089
  BT: human diseases
  BT: nutritional disorders
  NT: wasting disease
  RT: anaemia
  RT: malnutrition
```

Figure 1: Sample CABI record

Fig. 1 shows an example CABI record in the original formatting.

4 Experiments set up

We conducted a set of four experiments. Table 2 summarizes the parameters of our experiments. Notice for the fourth experiment we use S-Match because applying Min S-Match for flat structures brings no advantages. The following steps could be varied in the experiment.

First, a conversion from thesauri formats can be performed in a variety of ways. The most important parameters that influence the final result include: how to import relations, how to resolve ambiguities arising during conversion process and which knowledge base to use. We import only BT and NT (narrower term) relations for establishing a hierarchy of concepts. During the import we found a number of terms that have multiple broader terms. Such concepts could be placed in two (or more) places in the final hierarchy. Instead of removing BT relations until one remains, we left these terms under all their broader terms to increase matching chances.

Second, we can preserve the hierarchy of terms using BT and NT relations, or we can match term to term without considering the hierarchy (flat match).

Third, we can use different knowledge bases. We used two knowledge bases: WordNet version 2.1 and a 400.000 version of Stanford WordNet Project.

Fourth, we can choose between standard semantic matching and minimal semantic matching.

Fifth, the input sources were changed for the last two experiments mostly for technical reasons. According to experts, the structure and content of the 2009 version of Agrovoc has been improved a lot in comparison with the 2007 version. However, the 2009 version was not available during the first two experiments, but due to the amount of changes it was decided that it would be beneficial to proceed with a new version.

The matching consists of four steps: preprocessing (or concept at label computation), contextualization (or concept at node computation), element-level matching and structure-level matching. Below we will refer to some parameters and figures related to these stages of matching process.

Parameter	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Agrovoc version	2007-08-10	2007-08-10	2009-11-03	2009-11-03
CABI version	2009-11-01	2009-11-01	2009-11-01	2009-11-01
Agrovoc terms-leaves	35 036	35 036	47 805	40 574
CABI terms-leaves	29 172	29 172	29 172	24 241
Conversion	hierarchy	hierarchy	hierarchy	terms only
Knowledge base	WordNet 2.1	SWN 400K	SWN 400K	SWN 400K
Matching algorithm	MinSMatch	MinSMatch	MinSMatch	S-Match

Table 2: Experiments parameters

Table 3 summarizes the quantitative results of the preprocessing stage. Using a general-purpose knowledge base such as WordNet on a domain-specific input results in a high amount of unrecognized words. For these words the matcher has to rely only on string-based matching techniques. Using extended WordNet from Stanford WordNet Project results in slightly improved coverage. Differences in coverage also depend on the differences in thesauri versions and on the conversion parameters.

Parameter	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Knowledge base	WordNet 2.1	SWN 400K	SWN 400K	SWN 400K
Unrecognized words in Agrovoc	16 080	14 934	15 725	17 301
Unrecognized words in CABI	18 235	16 890	17 537	18 890

Table 3: Preprocessing stage figures

5 Experimental results

Table 4 summarizes the results of the experiments. Using extended knowledge base on the element-level matching step increases mapping size. Relatively small amount of equivalence relations should be noted in the first experiments. In the fourth experiment, where BT/NT relations were not used for conversion and only plain terms were matched, the amount of discovered equivalence links is significantly larger.

In the latter case the algorithm was able to establish an equivalence (EQ) relation directly between two terms, while in the former cases it failed to establish the relation when intermediate terms were present in the hierarchy. We hypothesize that if the pairs of terms in question are the same, then this could be due to the lack of background knowledge. That is, in the former cases, a proper relation was not established between the intermediate terms, thus preventing the establishment of a relation between the end terms. Another option is that this is a consequence of using minimal match algorithm. Namely, the relation was established from one term to another, but either remained as a derived one and to be found in the maximized mapping (unlikely, because the amount of EQs in maximized mapping is roughly the same), or again, lack of background knowledge prevented the establishment of a relation between intermediate terms, in turn preventing the establishment of a relation between the end terms.

We report here both maximized and minimized mapping sizes due to their different purposes. The minimized mapping contains sort of "compressed information", leaving out many links, which could be derived. Therefore it is useful for exploration and validation as it minimizes the effort required [14]. If used with applications, however, the consuming application should be aware of semantics of the minimal mapping. The maximized mapping has traditional semantics and is ready for the immediate consumption by applications. The difference between minimized and maximized mapping sizes reaches 17 times.

For the fourth experiment, mapping maximization (or minimization) is not applicable due to the absence of hierarchy. For convenience only, we report the results together with minimized mappings.

	Experiment 1		**Experiment 2**		**Experiment 3**		**Experiment 4**	
	min	max	min	max	min	max	min	max
EQ	3 698	3 564	3 603	3 468	3 009	3 407	28 457	N/A
DJ	125 439	3 811 923	124 648	3 777 493	184 564	8 304 315	199 836	N/A
MG	84 759	204 665	83 931	173 992	120 464	253 161	882 331	N/A
LG	218 579	1 262 700	223 140	123 6684	312 548	2 155 002	1 084 186	N/A
Total	432 475	5 282 852	435 322	5 191 637	620 585	10 715 885	2 194 810	N/A

Table 4: Experiments results

Table 5 provides the runtime and the memory as reported by the Java virtual machine during the experiments. The experiments were executed on a laptop with Intel Core 2 Duo T9600 processor and 4G RAM under Windows 7 x64 using the 64-bit Java machine. The run times should be considered approximate, because, although S-Match currently runs single-threaded and there were two processors available with one available almost exclusively for JVM, the matching process was not the only process in the OS and other (lightweight) activities were conducted during the experiments.

Experiment	1	2	3	4
Memory used, Mb	2 082	1 718	2 982	1 104
Run time, hours	˜10,5	˜12	˜27	˜7,5

Table 5: Experiments run time and memory

The significantly larger run time of the 3rd experiment could be explained by the fact that JVM was using all available memory and parts of it were swapped, slowing down calculations.

6 Evaluation

In matching experiments, evaluation is not a simple task [1]. For large matching tasks, such as this one, many of the more precise techniques based on a manual examination are not applicable due the size of the data. Other techniques for large data sets evaluation, such as reported in [10], exploit availability of instances to obtain the golden standard, and were not applicable in this case.

To evaluate the quality of links discovered by the matching algorithm, we need a golden standard to compare the mapping to. Such a mapping is usually created by an expert in the domain of the resources being matched and not only requires significant effort, but in many cases is impossible to create. Expert time is an extremely valuable resource and there is but a little of it available. This limits us in choosing an evaluation method.

We have chosen to evaluate a random sample of links from the mapping. We have used a sample of 200 randomly selected links. In the following we assume that the mapping being evaluated contains links with 4 relations: EQ (equivalence), DJ (disjointness), LG (less generality), MG (more generality). The part of the mapping consisting of EQ, LG and MG links is called "positive" part. The rest, namely DJ links, is called "negative" part.

Traditionally, the most interesting part of the mapping is the positive part, with equivalences being the most desired links. However, one should consider the value of the mapping together with its intended use, keeping the target application in mind. For example, traditionally DJ relations are discarded as being of non interest. However, if the mapping is used for search purposes, DJ relations could be used to prune search space and therefore shorten search times. Similar reasoning could be done with less or more general links for narrowing or broadening search in a manner similar to how BT/NT relations work.

6.1 Methodology

Non-trivial nature of evaluating matching algorithm leads us to splitting the evaluation into two phases. The first phase is a "relaxed" evaluation of the matching results. This is a commonly used approach and it does not take into account the relation returned by the matcher. Only the presence of the relation is considered and it is treated like a similarity between two terms. This approach is applicable only to the positive part of the mapping.

The second phase of the evaluation is a "strict" evaluation of the results. It does take a link relation into account. The evaluation of the matching results in both cases was conducted by a single expert, actively involved in the development of the thesauri.

Let us illustrate the relaxed and strict evaluation with the example. Consider the link in Figure 2. This link is considered by an expert as valid under relaxed conditions. Definitely, there is a relation between Egypt and Suez Canal. However, this relation is not *less generality*. Therefore, the link is considered invalid under strict evaluation conditions.

```
\Africa\North Africa\Egypt < \Middle East\Egypt\Suez Canal
```

Figure 2: Relaxed link example

We hypothesize that these kinds of approximations in link relations originate from various approximations in translation of semantics of input sources and semantics of knowledge bases, as well as approximations in source data sets and knowledge base data. For example, in the source data sets no explicit difference is made between partOf and isA relations. These two fundamental relations are often mixed in the same hierarchy, which leads to a less precise conversion to lightweight ontologies [12].

6.2 Results

This section presents the results of the evaluation. Given the resources available, it was not possible to evaluate recall, therefore, we report only precision. Moreover, the sample we have been able to evaluate is extremely small compared to the mapping size, therefore these figures should be considered only as an approximation. How close they approximate exact figures (obtained by evaluating the complete mapping) and how big should be the sample to have a fair approximation is still a research issue [1].

The following "facets" of the results are available. We differentiate between strict and relaxed evaluation, between overall, positive and negative parts of the mapping, between minimized and maximized mappings. Relaxed evaluation gives traditional precision measure and allows some degree of comparison with other matching systems. Strict evaluation gives a closer and more rigorous view of the results. Due to varying degree of interest of different mapping parts we provide overall, positive and negative parts precision separately. In addition to the reasons mentioned above, our recent report [1] gives more details on why we differentiate minimized and maximized mappings precision.

Table 6 and Table 7 show precision figures for the relaxed and strict evaluation scenarios, respectively. The overall line contains precision for the complete set of links. Positive and negative lines show precision for the positive and negative parts of the mapping, respectively. "min" columns stand for minimized mapping precision and "max" columns stand for maximized mapping precision.

Experiment	1		2		3	4	
	min	max	min	max	N.A	min	max
Positive	**18.60**	14.08	10.49	14.61	N.A	06.98	N.A
Negative	97.18	52.15	94.74	99.13	N.A	100.00	N.A
Overall	25.81	31.45	21.74	21.74	N.A	34.28	N.A

Table 6: Relaxed precision for minimized and maximized mappings, %

Experiment	1		2		3	4	
	min	max	min	max	N.A	min	max
Positive	**05.43**	03.30	02.80	01.38	N.A	03.49	N.A
Negative	97.18	52.15	94.74	99.13	N.A	100.00	N.A
Overall	38.00	21.41	29.00	35.28	N.A	14.87	N.A

Table 7: Strict precision for minimized and maximized mappings, %

In the fourth experiment a significantly larger amount (28 457) of equivalences was discovered and we decided to evaluate them separately, again, using strict and relaxed scenarios. To make it more interesting, we removed trivial cases (8 654) of equal strings out of this set and evaluate only a sample out of the remaining 19 782 non-trivial equivalence links. Here we obtain 27% and 16.50% precision in relaxed and strict scenarios, respectively. The mapping produced in this experiment was selected for the future processing and use by the users.

7 Conclusions

The experiment allowed us to accomplish several goals. First, the matching algorithm was tested on a domain-specific data, showing that further research and improvements in acquiring and using domain specific background knowledge is needed. Second, it was stress-tested on large data sets. As far as we know, this is the largest matching task tried. Third, the experiment provides one more confirmation for a trend observed in other cases: many algorithms show high precision and recall values on small "toy" data sets, with a decrease in performance with the increasing data set size. It therefore further confirms the need for a large and diverse golden standards for better evaluations of matching algorithms, including their robustness aspects. Last, we created several versions of the mappings, some of which were used by the users.

Precision figures reported here allow one to use created mappings accordingly. High precision parts (such as negative parts) could already be used without reservations. Positive parts with sufficiently high precision, such as the equivalences from the fourth experiment, can be used as a basis for further manual validation.

8 Future work

This experiment confirmed and further outlined the following directions of the future work. The experiment results can be improved by using the natural language processing pipeline to improve translation into logics. Currently, there are some phenomena in thesauri that remain unaddressed by the current heuristics and are translated incorrectly. These include round brackets and their use for disambiguation as well as use of comma for term specification and qualification.

Another promising direction is more carefully importing knowledge from existing sources to augment current knowledge base with more relations and terms, achieving better domain coverage and domain specificity. This includes the knowledge from input sources (CABI and Agrovoc), as well as other thesauri covering the domain, such as the National Agricultural Library Thesaurus (NALT).

To analyse further the performance of the algorithm and ways to improve it, an extended analysis of current results would be beneficial. There are two ways to proceed with this task. First, for the first time we have a strict evaluation together with a relaxed one and we can analyse the reasons why the relation is not established correctly. Second, a set of false positives (links discovered, but judged to be incorrect) can be analysed to discover the errors in matching and ways to fix them. Third, a set of correct links (need to be established by an expert) would allow calculating recall and analysing the reason why the algorithm misses the links.

References

[1] Aliaksandr Autayeu, Vincenzo Maltese, and Pierre Andrews. Recommendations for qualitative ontology matching evaluations. In Shvaiko et al. [17].

[2] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *SIGMOD Record*, 35(3):34–41, 2006.

[3] Anhai Doan and Alon Y. Halevy. Semantic integration research in the database community: A brief survey. *AI Magazine*, 26:83–94, 2005.

[4] Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer-Verlag New York, Inc., 2007.

[5] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[6] Fausto Giunchiglia, Aliaksandr Autayeu, and Juan Pane. S-Match: an open source framework for matching lightweight ontologies. *Semantic Web Journal*, 2010.

[7] Fausto Giunchiglia, Vincenzo Maltese, and Aliaksandr Autayeu. Computing minimal mappings. In Shvaiko et al. [17].

[8] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Discovering missing background knowledge in ontology matching. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *ECAI*, volume 141 of *Frontiers in Artificial Intelligence and Applications*, pages 382–386. IOS Press, 2006.

[9] Fausto Giunchiglia, Dagobert Soergel, Vincenzo Maltese, and Alessandro Bertacco. Mapping large-scale knowledge organization systems. In *ICSD*, 2009.

[10] Fausto Giunchiglia, Mikalai Yatskevich, Paolo Avesani, and Pavel Shvaiko. A large dataset for the evaluation of ontology matching systems. *The Knowledge Engineering Review Journal*, 24(2):137–157, 2009.

[11] Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic matching: Algorithms and implementation. *J. Data Semantics*, 9:1–38, 2007.

[12] Fausto Giunchiglia and Ilya Zaihrayeu. Lightweight ontologies. In M. Tamer Ozsu and Ling Liu, editors, *EoDS*, pages 1613–1619. Springer, 2009.

[13] Boris Lauser, Gudrun Johannsen, Caterina Caracciolo, Johannes Keizer, Willem Robert van Hage, and Philipp Mayr. Comparing human and automatic thesaurus mapping approaches in the agricultural domain. *CoRR*, abs/0808.2246, 2008.

[14] Vincenzo Maltese, Fausto Giunchiglia, and Aliaksandr Autayeu. Save up to 99% of your time in mapping validation. In Robert Meersman, Tharam S. Dillon, and Pilar Herrero, editors, *OTM Conferences (2)*, volume 6427 of *Lecture Notes in Computer Science*, pages 1044–1060. Springer, 2010.

[15] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. *JoDS*, 4:146–171, 2005.

[16] Pavel Shvaiko and Jérôme Euzenat. Ten challenges for ontology matching. In Robert Meersman and Zahir Tari, editors, *OTM Conferences (2)*, volume 5332 of *Lecture Notes in Computer Science*, pages 1164–1182. Springer, 2008.

[17] Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Natalya Fridman Noy, and Arnon Rosenthal, editors. *Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009*, volume 551 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[18] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *ACL*. The Association for Computer Linguistics, 2006.

[19] Ilya Zaihrayeu, Lei Sun, Fausto Giunchiglia, Wei Pan, Qi Ju, Mingmin Chi, and Xuanjing Huang. From web directories to ontologies: Natural language processing challenges. In *ISWC/ASWC*, pages 623–636, 2007.