Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System

Roberto NAVIGLI, Paola VELARDI
Dipartimento di Informatica
Università "La Sapienza"
via Salaria 113
Roma, Italy, 00198
{velardi,navigli}@di.uniroma1.it

Alessandro CUCCHIARELLI, Francesca NERI
DIIGA
Università Politecnica delle Marche
via Brecce Bianche 12
Ancona, Italy, 60131
{cucchiarelli, neri}@diiga.univpm.it

Abstract

Ontology evaluation is a critical task, even more so when the ontology is the output of an automatic system, rather than the result of a conceptualisation effort produced by a team of domain specialists and knowledge engineers. This paper provides an evaluation of the OntoLearn ontology learning system. The proposed evaluation strategy is twofold: first, we provide a detailed *quantitative* analysis of the ontology learning algorithms, in order to compute the accuracy of OntoLearn under different learning circumstances. Second, we automatically generate natural language descriptions of formal concept specifications, in order to facilitate per-concept *qualitative* analysis by domain specialists.

1 Evaluating ontologies

Automatic methods for ontology learning and population have been proposed in recent literature (e.g. ECAI-2002 and KCAP-2003 workshops[1]) but a co-related issue then becomes the *evaluation* of such automatically generated ontologies, not only with the goal of comparing the different approaches (Hovy, 2001) and ontology-based tools (Angele and Sure, 2002), but also to verify whether an automatic process may actually compete with the typically human process of converging on an *agreed* conceptualization of a given domain. Ontology construction, apart from the technical aspects of a knowledge representation task (i.e. choice of representation languages, consistency and correctness with respect to axioms, etc.), is a *consensus building* process, one that implies long and often harsh discussions among the specialists of a given domain. Can an automatic method simulate this process? Can we provide domain specialists with a means to measure the *adequacy* of a specific set of concepts as a model of a given domain?, Specialists are often unable to evaluate the *formal content* of a computational ontology (e.g. the denotational theory, the formal notation, the knowledge representation system capabilities like property inheritance, consistency, etc.). Evaluation of the formal content is rather tackled by computational scientists, or by automatic verification systems. The role of the specialists is instead to compare their intuition of a domain with the description of this domain, as provided by the ontology concepts. To facilitate one such *qualitative* per-concept evaluation, we devised a method for automatic generation of textual explanations (*glosses*) of automatically learned concepts. Glosses provide a description, in natural language, of the formal specifications assigned to the learned concepts. An expert can easily compare his intuition with these natural language descriptions.

The objective of the gloss-based evaluation is, as previously remarked, to obtain a judgement, by domain specialists, concerning the adequacy of an automatically derived domain conceptualisation. On the computational side, an ontology learning tool is based on a battery of software programs aimed at extracting and formalising domain knowledge, usually starting from unstructured data. Therefore, it is equally important to produce a detailed evaluation of these programs, on a *quantitative* ground, in order to gain insight on the internal and external contingencies that may affect the result of an ontology learning process.

In what follows, we firstly provide a quantitative evaluation of the OntoLearn ontology learning system, under different learning circumstances. Secondly, we describe the gloss-based per-concept evaluation method. Both evaluation strategies are experimented in two application domains: Tourism and Economy.

The subsequent section provides a sketchy description of the OntoLearn algorithms. Details are found in (Navigli and Velardi, 2004) and (Navigli, Velardi and Gangemi, 2003). Sections 3

[1] ECAI-2002 http://www-sop.inria.fr/acacia/WORKSHOPS/ECAI2002-OLT/accepted-papers.html
KCAP-2003 http://km.aifb.uni-karlsruhe.de/ws/semannot2003/papers.html

and 4 are dedicated to the quantitative and qualitative analyses of OntoLearn.

2 Summary of the OntoLearn system

OntoLearn is an ontology population method based on text mining and machine learning techniques. OntoLearn starts with an existing *generic ontology* (we use WordNet, though other choices are possible) and a set of documents in a given domain, and produces a domain extended and trimmed version of the initial ontology. The ontology generated by OntoLearn is anchored to texts, it can be therefore classified as a *linguistic ontology* (Gòmez-Pérez et al. 2004).

OntoLearn has been applied to different domains (tourism, computer networks, economy) and in several European projects[2].

Concept learning is achieved in the following three phases:

1) **Terminology Extraction**: A list of domain multi-word expressions (MWE hereafter) is extracted from a set of documents that are judged representative of a given domain. MWEs are extracted using natural language processing and statistical techniques. Contrastive corpora and glossaries in different domains are used to prune terminology that is not domain-specific. Domain MWEs are selected also on the basis of an entropy-based measure that *simulates specialist consensus* on concepts choice: in words, the probability distribution of a "good" domain MWE must be uniform across the individual documents of the domain corpus.
2) **Semantic interpretation of MWEs**: Semantic interpretation is based on a principle, *compositional interpretation*, and on a novel algorithm, called *structural semantic interconnections* (SSI). Compositional interpretation signifies that the meaning of a multi-word expression (MWE) can be derived compositionally from its components[3], e.g. the meaning of *business plan* is derived first, by associating the appropriate concept identifier, with reference to the initial top ontology, to the component terms (i.e. sense #2 of *business* and sense #1 of *plan* in WordNet), and then, by identifying the semantic relations holding among the involved concepts (e.g. $plan\#1 \xrightarrow{topic} business\#2$).
3) **Extending and trimming the initial ontology**: Once the terms have been semantically interpreted, they are organized in sub-trees, and appended under the appropriate node of the initial ontology, e.g. $business\_plan\#1 \xrightarrow{kind\_of} plan\#1$. Furthermore, certain upper and lower nodes of the initial ontology are pruned to create a *domain-view* of the ontology. The final ontology is output in OWL language.

SSI lies in the area of *syntactic pattern matching* algorithms (Bunke and Sanfeliu, 1990). It is a word sense disambiguation algorithm used to determine the correct sense (with reference to the initial ontology) for each component of a complex MWE. The algorithm is based on building a graph representation for alternative senses of each MWE component[4], and then selecting the appropriate senses on the basis of detected *semantic interconnection patterns* between graph pairs. The SSI algorithm seeks for semantic interconnections among the words of a *context* T. Contexts T_i are generated from groups of partially overlapping complex MWEs (extracted during phase 1 of the OntoLearn procedure) sharing the same *syntactic head*. For example, given the list of complex MWEs *securities portfolio, investment portfolio, real-estate portfolio, junk-bond portfolio, diversified portfolio, stock portfolio, bond portfolio, loan portfolio,* the following list of term components is created:

T=[*security, investment, real-estate, estate, bond, junk-bond, diversified, stock, portfolio, loan*]

Relevant pattern *types* are described by a context free grammar G. An example of rule in G is the following (S_1 S_2 and S are concepts, i.e. synsets in WordNet):

Rule Name:*gloss+hyperonymy/meronymy* (S_1,S_2):

Def: $\exists G \in Synsets : S_1 \xrightarrow{gloss} S$ and there is a hyperonymy/meronymy path between S and S_2

For instance, in *railways company*, the gloss of *railway#1* contains the word *organization*, and there is an hyperonymy path of length 2 between company#1 and organization#1. That is: $railway\#1 \xrightarrow{gloss} organization\#1$, and: $company\#1 \xrightarrow{kind\_of} institution\#1 \xrightarrow{kind\_of} organization\#1$. This pattern (an instance of the *gloss+hypeonymyr/meronymy* rule) cumulates

[2] E.g. : Harmonize IST-2000-29329 and the INTEROP network of excellence, started on december 2003.

[3] In the literature, multi word expressions are classified as compositional, idiosyncratically compositional and non-compositional. In mid-technical domains, compositional MWEs cover about 60-70% of MWE (we cannot support with data our statitics for sake of space)

[4] We remark again that a detailed description of the SSI algorithm is in (Navigli & Velardi, 2004) and (Navigli, Velardi and Gangemi, 2003). Graphs are generated on the basis of lexico-semantic information in WordNet and in a variety of on-line resources, see the mentioned papers for details.

evidence for senses #1 of both *railway* and *company*.

In SSI, the *correct sense* S_t for a term t∈ T is selected depending upon the number and weight of patterns matching with rules in G. The weights of patterns are automatically learned using a perceptron[5] model. The weight function is given by:

$$(1) weight(pattern_j) = \alpha_j + \beta_j(\frac{1}{length\_pattern_j})$$

where α_j is the weight of rule j in G, and the second addend is a smoothing parameter inversely proportional to the length of the matching pattern (e.g. 2 in the previous example, since 2 is the minimal length of the rule, and the actual length of the pattern is 3). The perceptron has been trained on the SemCor[6] semantically annotated corpus.

In order to complete the semantic interpretation process, OntoLearn then attempts to determine the *semantic relations* that hold between the components of a complex concept. In order to do this, it was first necessary to select an inventory of semantic relations. We examined several proposals, like EuroWordnet (Vossen, 1999), DOLCE (Masolo et al., 2002), FrameNet (Ruppenhofer Fillmore & Baker, 2002) and others.

As also remarked in (Hovy, 2001), no systematic methods are available in literature to compare the different sets of relations. Since our objective was to define an automatic method for semantic relation extraction, our final choice was to use a reduced set of FrameNet relations, which seemed general enough to cover our application domains (tourism, computer networks, economy). The choice of FrameNet is motivated by the availability of a sufficiently large set of annotated examples of conceptual relations[7], that we used to train an available machine learning algorithm, TiMBL (Daelemans et al., 2002). The relations used are: *Material, Purpose, Use, Topic, Product, Constituent Parts, Attribute*[8]. Examples for each relation are the following:

$net\#1 \xleftarrow{attribute} loss\#3$

$takeover\#2 \xleftarrow{topic} proposal\#1$

$sand\#1 \xleftarrow{material} beach\#1$

$merger\#1 \xleftarrow{purpose} agreement\#1$

$meeting\#1 \xleftarrow{use} room\#1$

$bond\#2 \xleftarrow{const\_part} market\#1$

$computer\#1 \xleftarrow{product} company\#1$

We represented training instances as pairs of concepts annotated with the appropriate conceptual relation, e.g.:

`[(computer#1,maker#3),Product]`

Each concept is in turn represented by a *feature-vector* where attributes are the concept's hyperonyms in WordNet, e.g.:

```
(computer#1,maker#3):
      ((computer#1,machine#1,device#1
,instrumentality#3),(maker#3,business
#1,enterprise#2,organization#1))
```

3 Quantitative Evaluation of OntoLearn

This section provides a quantitative evaluation of OntoLearn s main algorithms. We believe that a quantitative evaluation is particularly important in complex learning systems, where errors can be produced at almost any stage. Even though some of these errors (e.g. subtle sense distinctions) may not have a percievable effect on the final ontology, as shown by the results of the qualitative evaluation in Section 4.2, it is nevertheless important to gain insight on the actual system capabilities, as well as on the pararmeters and external circumstances that may positively or negatively influence the final performance.

3.1 Evaluating the MWE extraction algorithm

The terminology extraction algorithm has been evaluated in the context of the European project Harmonise on Tourism interoperability. We first collected a corpus of about 1 million words of tourism documents, mainly descriptions of travel and tourism sites. From this corpus, a syntactic parser extracted an initial list of 14,383 candidate complex MWEs from which the statistical filters selected a list of 3,840 domain-relevant complex MWEs, that were submitted to the domain specialists. The Harmonise ontology partners were not skilled to evaluate the OntoLearn semantic interpretation of MWEs, therefore we let them evaluate only the domain appropriateness of the *terms*. The gloss generation method described in Section 4 was subsequently concieved to overcome this limitation.

We obtained a precision ranging from 72.9% to about 80% and a recall of 52.74%. The precision shift is due to the well-known fact that experts may have different intuitions about the relevance of a concept. The recall estimate was produced by

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://www.cs.unt.edu/~rada/downloads.html#semcor

[7] The choice of FrameNet was motivated more by availability than appropriateness.

[8] The relation Attribute is not in FrameNet, however it was a useful relation for terminological strings of the adjective_noun type.

manually inspecting 6,000 of the initial 14,383 candidate MWEs, asking the experts to mark all the MWEs judged as "good" domain MWEs, and comparing the obtained list with the list of terms automatically filtered by OntoLearn.

We ran similar experiments on an Economy corpus and a Computer Network corpus, but in this case the evaluation was performed by the authors. Overall, the performance of the MWE extraction task appears to be influenced by the dimension and the *focus* of the starting corpus (e.g. "generic tourism" vs. "hotel accomodation descriptions"). Small and unfocused corpora do not favor the efficacy of statistical analysis. However, the availability of sufficiently large and focused corpora seems a realistic requirement for most applications.

3.2 Evaluating the ontology learning algorithms

The distinctive task performed by OntoLearn is semantic disambiguation. The performance of the SSI algorithm critically depends upon two factors: the first is the ability to detect *semantic interrelations* among concepts associated to the words of complex MWEs, the second is the *dimension of the context* T available to start the disambiguation process.

As for the first factor, there are two possible ways of enhancing reliable identification of semantic interconnections: one is to tune at best the weight of individual rules in G (e.g. formula (1) in Section 2), the second is to enrich the semantic information associated to alternative word senses. The latter is an on-going research activity.

As far as the context T is concerned, the intuition is that, with a larger $|T|$, there are higher chances of detecting semantic patterns among the "correct" senses of the terms in T. However, the dimension of contexts T_i is an external contingency, it depends upon the available corpus.

Accordingly, we evaluated the SSI algorithm using as parameters the dimension of T, $|T|$, and the weights associated to rules in G. We ran several experiments over the full terminology extracted from the Economy and Tourism corpora, but performances are computed only on, respectively, 453 and 638 manually disambiguated terms. This means that in a context T_i including, e.g. k terms, we evaluate OntoLearn's sense choices only for the fragment of $j \leq k$ terms, for which the "true" sense has been manually assigned.

Table 1 shows the performance of SSI (precision and recall) when using only patterns whose weight, computed with formula (1) is over a threshold ϑ. The "Core" column in Table 1 shows the performance of SSI when accepting *only* these core patterns, while the third column refers to all matching patterns. With $\vartheta = 0,7$ a subset of 7-9 rules[9] in G (over a total of 20) are used by the algorithm. Interestingly enough, these rules have a high probability of being hired, as shown by the relatively low difference in recall. The Baseline tower in Table 1 is computed selecting always the first sense (senses in WordNet are ordered by probability in everyday language).

Table 2 shows that performance of SSI is indeed affected by the dimension of T. Large $|T|$, as expected, improves the performance, however, overly large contexts (>80 terms) may favor the detection of non-relevant patterns.

In general, both experiments show that the Economy corpus performs better than the Tourism, since the latter is less technical (the baseline is quite high), rather unfocused, and contexts T_i are much less populated.

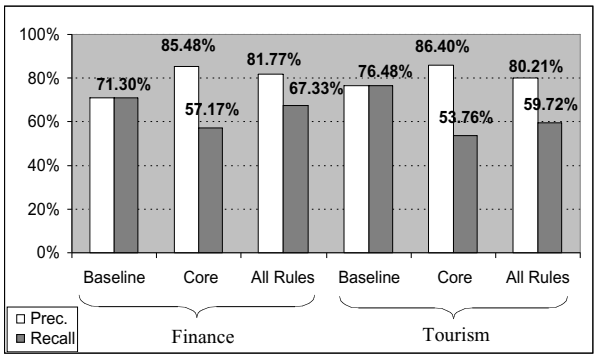

Table 1. Performances as a function of pattern's weight

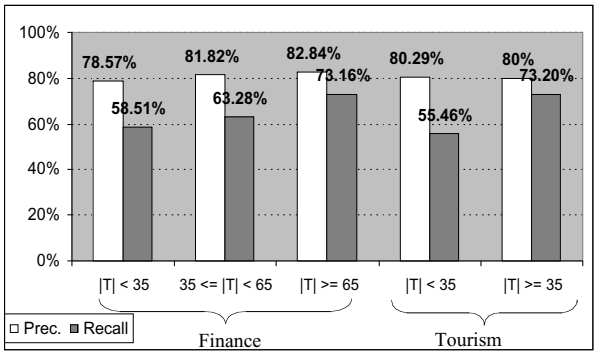

Table 2. Performances as a function of |T|

We remark that SSI performs better than standard WSD (word sense disambiguation) tasks but this is also motivated by the fact that context words in T are more interrelated than co-occurring words in generic sentences. The SSI algorithm, by

[9] in formula (1), α, that depends upon the rule, has a much higher influence than β, that depends upon the matching pattern)

its very nature, is favored by focused and large contexts. In any case, it is worth mentioning that SSI received the second best score in the latest SenSeval-3[10], gloss disambiguation exercise, placed about 1% below the first and about 11% before the third participant.

3.3 Evaluating the semantic annotation algorithm

To test the semantic relation annotation task, we used a learning set (including selected annotated examples from FrameNet (FN), Tourism (Tour), and Economy (Econ)), and a test set with a distribution of examples shown in Table 3.

Table 3. Distribution of examples in the learning and test set for the semantic annotation task

	Learning Set				Test Set			
Sem_Rel	**FN**	**Tour**	**Econ**	**Tot**	**FN**	**Tour**	**Econ**	**Tot**
MATERIAL	8	3	0	*11*	5	2	0	*7*
USE	9	32	2	*43*	6	20	1	*27*
TOPIC	52	79	100	*231*	29	43	50	*122*
C_PART	3	7	12	*22*	2	4	6	*12*
PURPOSE	26	64	22	*112*	14	34	11	*59*
PRODUCT	3	1	6	*10*	1	1	4	*6*
Total	101	186	142	429	57	104	72	233

Notice that the relation *Attribute* is generated whenever the term associated to one of the concepts is an adjective. Therefore, this semantic relation is not included in the evaluation experiment, since it would artificially increase performances. We then tested the learner on test sets for individual domains[11], leading to the results shown in Table 4 a and b.

Table 4 Performance of the semantic annotation task on a) Tourism b) Economy

	d<=10%	**d<=30%**	**d<=100%**
Precision MACRO	0,958	0,875	0,847
Recall MACRO	0,283	0,636	0,793
F1 MACRO	0,437	0,737	0,819
Precision micro	0,900	0,857	0,798
Recall micro	0,087	0,635	0,798
F1 micro	0,158	0,721	0,798

	d<=10%	**d<=30%**	**d<=100%**
Precision MACRO	1,000	0,804	0,651
Recall MACRO	0,015	0,403	0,455
F1 MACRO	0,030	0,537	0,536
Precision micro	1,000	0,758	0,750
Recall micro	0,042	0,653	0,750
F1 micro	0,080	0,701	0,750

The performance measures are those adopted in TREC competitions[12]. The parameter **d** in the above Tables is a confidence factor defined in the TiMBL algorithm. This parameter can be used to increase system's robustness in the following way: whenever the confidence associated by TiMBL to the classification of a new instance is lower than a given threshold, we output a "generic" conceptual relation, named *Relatedness.* We experimentally fixed the threshold for **d** around 30% (central column of Table 4).

Table 4 demonstrates rather good performances, however the main problem with semantic relation annotation is the unavailability of an agreed set of conceptual relations, and a sufficiently large and balanced training set. Consequently, we need to update the set of used relations whenever we analyse a new domain, and re-run the training phase enriching the training corpus with manually tagged examples from the new domain (as for in Table 2).

4 Qualitative evaluation: Evaluating the generated ontology on a per-concept basis

The lesson learned during the Harmonise EC project was that the domain specialists, tourism operators in our case, can hardly evaluate the formal aspects of a computational ontology. When presented with the domain extended and trimmed version of WordNet (OntoLearn's phase 3 in Section 2), they were only able to express a generic judgment on each node of the hierarchy, based on the concept label. These judgments were used to evaluate the terminology extraction task, but the experiment suggested that, indeed, it was necessary to provide a better description for the learned concepts.

4.1 Gloss generation grammar

To help human evaluation on a per-concept basis, we decided to enhance OntoLearn with a gloss generation algorithm. The idea is to generate glosses in a way that closely reflects the key aspects of the concept learning process, i.e. semantic disambiguation and annotation with a conceptual relation.

The gloss generation algorithm is based on the definition of a grammar with distinct generation rules for each type of semantic relation.

Let $s_i^h \xrightarrow{sem\_rel} s_j^k$ be the complex concept associated to a complex term $w_h w_k$ (e.g. *jazz festival*, or *long-term debt*), and let:

<H>= the syntactic head of $w_h w_k$ (e.g. *festival*, *debt*)

<M> = the syntactic modifier of $w_h w_k$ (e.g. *jazz*, *long-term*)

<GNC>= be the gloss of the new complex concept S^{hk}

<HYP>= the selected sense of <H>(e.g. respectively, *festival#1* and *debt#1*).

[10] SensEval–3 http://www.senseval.org/senseval3

[11] This of course penalised the results (the performance over a test set composed by examples of all the three domains is much higher), but provides a more realistic test bed of the generality of the approach.

[12] http://trec.nist.gov/

<MSGHYP>= the main sentence[13] of the WordNet gloss of <HYP>
<MSGM>= the main sentence of the WordNet gloss of the selected sense for <M>
Here we provide two examples of rules for generating GNCs:
If sem_rel=*Topic*, <GNC>:: = **a kind of** <HYP>**,** <MSGHYP>**, relating to the** <M>**,** <MSGM>**.**

e,g.: GNC(*jazz festival*): **a kind of** festival**,** a day or period of time set aside for feasting and celebration**, relating to the** jazz**,** a style of dance music popular in the 1920.
If sem_rel=*Attribute*, <GNC>:= **a kind of** <HYP>**,** <MSGHYP>**,** <MSGM>**.**

e.g.:GNC(*long term debt*)= **a kind of** debt, the state of owing something (especially money), relating to or extending over a relatively long time.

4.2 Per-concept evaluation experiment

To verify the utility of gloss generation, the automatically generated glosses were submitted for evaluation to two human experts, a tourism specialist from ECCA[14], and an economist from the University of Ancona. The specialists were not aware of the method used to generate glosses; they have been presented with a list of concept-gloss pairs and asked to fill in an evaluation form (see Appendix) as follows: vote 1 means "unsatisfactory definition", vote 2 means "the definition is helpful", vote 3 means "the definition is fully acceptable". Whenever he was not fully happy with a definition (vote 2 or 1), the specialist was asked to provide a brief explanation. For comparison, Appendix 2 shows also glossary definitions extracted from the web for the same MWEs, that were not shown to the specialists.

Table 5 provides a summary of the evaluation..

Table 5. Evaluation of glosses by domain specialists.

	vote =1	vote=2	vote=3	uncertain	average
Tourism total (97)	33 (34.7)	14 (14.4)	45 (46.3)	5 (5.1)	2,13
Ecomomy total (134)	52 (38.8)	16 (11.9)	66 (49.2)	-	2.10

The following conclusions can be drawn from this experiment:

1. Overall, the two domain specialists fully accepted the system's choices in 45-49% of the cases, and were reasonably satisfied in 12-14% of the cases. The average vote is above 2 in both cases.
2. As expected, if a MWE is compositional, the generated definition is more often accepted or fully accepted (e.g. examples 25_E and 14_T in Appendix 2). When a compositional interpretation is not accepted (vote=1), this is motivated either by an OntoLearn interpretation error (wrong sense or wrong conceptual relations) or by the unavailability of a correct sense in WordNet, despite the fact that the sense is not idiosyncratic. OntoLearn errors for compositional MWEs are 7 (5%) in Economy and 12 (13%) in Tourism. Examples of OntoLearn errors and core ontology "misses" are the definitions 14_T (wrong sense of *form*) and 19_E (no good sense for *bilateral* in WordNet), respectively.
3. Sometimes the specialists found it acceptable also an idiosyncratic or non compositional definition. This happens in 16 cases for the Tourism domain (16%) and in 19 cases for the Economy domain (13%). Examples are the MWEs 45_E and 76_E, both idiosyncratically decomposable, in Appendix 2.

One of the specialists is particularly involved in ontology building projects, therefore we report his valuable comment: "*some of the descriptions would not be appropriate to take them over in a tourism ontology just as they are. But most of them are quite helpful as basis for building the ontology. The most important problem from my point of view is the too detailed descriptions of the components itself instead of the meaning of the overall term in this context. Best example is the term "bed tax". Nobody would expect a definition of a bed or a tax.*" In other terms, he found disturbing the fact that a definition extensively reports the definitions of its components. On the other side, our objective is not only to produce concept definitions, but also to organize concepts in hierarchies. Showing the definitions of individual components is a "natural" mean to verify that the correct senses have been selected (e.g. the correct senses of *bed* and *tax*). This is clearly the case, since, for example in definition 14_T (*booking form*), the specialist was immediately able to diagnose a sense disambiguation error for *form*, though he was unaware of the OntoLearn methodology.

5 Concluding remarks

This paper presented an in-depth evaluation of the Ontolearn ontology learning system. The three basic algorithms (terminology extraction, sense disambiguation and annotation with semantic relation) have been individually evaluated in two

[13] The main sentence is the gloss pruned of subordinates, examples, etc.

[14] ECCA – eTourism Competence Center Austria.

domains, under different parametrizations, to obtain a realistic and comprehensible picture of system's capabilities. The critical algorithm, SSI, has very good performances that are favored by the fact that word sense disambiguation is applied to group of words (domain MWEs) that are strongly semantically related, unlike for generic WSD tasks (e.g. Senseval). The performance of the SSI algorithm can be further improved through an extension of the grammar G, which is an on-going research activity.

6 Acknowledgements

Our thanks go to Dr. Wolfram Höpken, from ECCA – eTourism Competence Center Austria (wolfram@hoepken.org) and Dr. Renato Iacobucci, from the University of Ancona, who gave up their precious time to evaluate our glosses. This work has been in part supported by the INTEROP Network of Excellence IST-2003-508011

References

J. Angele and Y. Sure (2002) "Whitepaper: Evaluation of Ontology-based Tools", Workshop on evaluation of ontology-based tools (EON2002), at the 13th Int. EKAW 2002, Sigueza (Spain), September 2002.

H. Bunke and A. Sanfeliu (editors) (1990). Syntactic and Structural pattern Recognition: Theory and Applications, World Scientific, Series in Computer Science vol. 7, 1990.

Daelemans,W. Zavrel, J. Van den Sloot, K. & Van den Bosch, A. (2002). TiMBL: Tilburg Memory Based Learner. Version 4.3 Reference Guide. Tilburg University.

Gòmez-Pérez, A., Fernàndez-Lopez M. and Corcho O. (2004). Ontological Engineering, Springer Verlag, London, 2004.

Hovy, E. (2001). Comparing Sets of Semantic relations in Ontologies. In R. Geen, C.A. Bean and S. Myaeng Semantic of relations. Kluwer.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. Oltramari, A. & Schneider, L. (2002). Sweetening Ontologies with DOLCE. Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web.

Navigli, R. & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. Computational Linguistics, MIT press, (50)2.

Navigli, R., Velardi, P. Gangemi, A. (2003). Corpus Driven Ontology Learning: a Method and its Application to Automated Terminology Translation. IEEE Intelligent Systems (18)1.22-31.

Ruppenhofer, J., Fillmore, C.J. & Baker, C.F. (2002). Collocational Information in the FrameNet Database. In Braasch, A. and Povlsen, C. (eds.), Proceedings of the Tenth Euralex International Congress. Copenhagen, Denmark. Vol. I: 359--369, 2002.

Vossen, P. (1999). EuroWordNet: General Document. Version 3 Final. http://www.hum.uva.nl/~ewn

APPENDIX: Excerpt of the per-concept evaluation form

Concept #: 25_E	**Term:** *business_plan*	**Synt:** N-N	**Rel<w_1,w_2>:** Topic
Gloss: a kind of plan, a series of steps to be carried out or goals to be accomplished, relating to the business, the activity of providing goods and services involving financial and commercial and industrial aspects.			
Specialist vote: 3			
Comment by Specialist: none			
Diagnose: none			
Glossary definition: a written report that states what a company (or a part of a company) aims to do increase sales, develop new products, etc. within a certain period, and how it will obtain the necessary finances and resources.			

Concept #: 2_T	**Term:** *affiliated_hotel*	**Synt:** Agg-N	**Rel<w_1,w_2>:** Attribute
Gloss: a kind of hotel, a building where travellers can pay for lodging and meals and other services, being joined in close association.			
Specialist vote: 3			
Comment by Specialist: none			
Diagnose: none			
Glossary definition: a hotel that is a member of a chain, franchise, or referral system. Membership provides special advantages, particularly a national reservation system.			

Concept #: 14_T	**Term:** *booking_form*	**Synt:** N-N	**Rel<w_1,w_2>:** Purpose
Gloss: a kind of form, alternative names for the body of a human being, for booking, the act of reserving (a place or passage) or engaging the services of (a person or group).			
Specialist vote: 1			
Comment by Specialist: definition of form wrong in this context			
Diagnose: OntoLearn disambiguation error for form			
Glossary definition: a document which purchasers of tours must complete to give the operator full particulars about who is buying the tour.			

Concept #: 19_E	**Term:** *bilateral_aid*	**Synt:** Agg-N	**Rel<w_1,w_2>:** Attribute
Gloss: a kind of aid, the activity of contributing to the fulfillment of a need or furtherance of an effort or purpose, having identical parts on each side of an axis.			
Specialist vote: 1			
Comment by Specialist: Fully wrong definition.			
Diagnose: WordNet gloss of bilateral is not adequate to domain (no better definition is available in WordNet).			
Glossary definition: assistance given by one country to another.			

Concept #: 45_E	**Term:** *cyclical_uneployment*	**Synt:** Agg-N	**Rel<w_1,w_2>:** Attribute
Gloss: a kind of unemployment, the state of being unemployed or not having a job, recurring in cycles.			
Specialist vote: 3			
Comment by Specialist: none			
Diagnose: none			
Glossary definition: workers are without a job because of a lack of aggregate demand due to a down turn in economic activity.			

Concept #: 76_E	**Term:** *foreign_aid*	**Synt:** Agg-N	**Rel<w_1,w_2>:** Attribute
Gloss: a kind of aid, the activity of contributing to the fulfillment of a need or furtherance of an effort or purpose, of concern to or concerning the affairs of other nations .			
Specialist vote: 3			
Comment by Specialist: none			
Diagnose: none			
Glossary defonition: the international transfer of public and private funds in the form of loans or grants from donor countries to recipient countries.			