TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities

Francesco Sclano, Paola Velardi

Dipartimento di Informatica, via Salaria 113 Roma
velardi@di.uniroma1.it, francesco_sclano@yahoo.it

Abstract

We implemented a high-performing technique to automatically extract from the available documents in a given domain the shared terminology of a web community. This technique has been successfully experimented and submitted for large-scale evaluation in the domain of enterprise interoperability, by the member of the INTEROP network of excellence. In order to make the technique available to any web community in any domain, we developed a web application that allows users to i) acquire (incrementally or in a single step) a terminology in any domain, by submitting documents of variable length and format, and ii) validate on-line the obtained results. The system also supports collaborative evaluation by a group of experts. The web application has been widely tested in several domains by many international institutions that volunteered for this task.

1. Introduction

In (Navigli and Velardi, 2004) we presented a technique, named OntoLearn, to automatically learn a domain ontology from the documents shared by the members of a web community. This technique is based on three learning steps, each followed by manual validation: terminology extraction, glossary extraction, and finally, ontology enrichment. The OntoLearn methodology has been enhanced (Navigli and Velardi, 2005) and experimented in real settings (Velardi et al. 2007). Recently, we started to develop web applications to make freely available each of the steps of the OntoLearn methodology. This paper describes TermExtractor, a tool to extract the terminology "shared" among the members of a web community, through the analysis of the documents they exchange. Defining a domain lexicon is in fact the first step of an ontology building process.

The contributions of the paper, with respect to published work, are the following:

1. We summarize the terminology extraction algorithm, on which we provided already a description in (Navigli and Velardi, 2002)[1], with the main intent of highlighting progress wrt previously reported work;
2. We summarize the features and options of the TermExtractor web application;
3. We provide an evaluation of the web tool performed by a world-wide group of TermExtractor users, who volunteered to perform the task.

[1] This paper intentionally focuses on the web application and evaluation more than on term extraction algorithms. The reader is invited to read the referred papers for additional details.

The main novelty is the evaluation procedure, considerably more wide-coverage than in available literature, where the standard methodology to evaluate terminology extraction systems is based on three judges with adjudication, on a single domain. Second, TermExtractor is able to learn the terminology from a corpus of documents, not just a single document. Finally, to the best of our knowledge, the TermExtractor web application is the most sophisticated freely available terminology-extraction tool to date.

2. The term extraction algorithms

As many terminology extraction systems (Wermter and Hahn, 2005) (Bourigault and Jacquemin, 1999) (Park et al., 2002), in TermExtractor the identification of relevant terms is based on two steps: first, a linguistic processor is used to parse text and extract typical terminological structures, like compounds (*enterprise model*), adjective-noun (*local network*) and noun-preposition-noun (*board of directors*). Then, the (usually large) list of terminological candidates is purged according to various filters.

Figure 1 shows the main processing phases of the term extraction module.

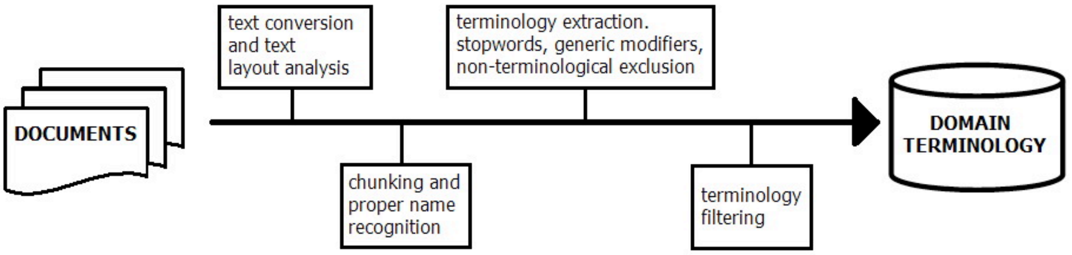

DOCUMENTS
text conversion and text layout analysis
chunking and proper name recognition
terminology extraction. stopwords, generic modifiers, non-terminological exclusion
terminology filtering
DOMAIN TERMINOLOGY

Figure 1. Terminology extraction phases.

TermExtractor combines the best reported terminology extraction techniques, including those introduced in (Navigli and Velardi, 2002). We summarize here the filters used to exclude non-terminological multi-word strings from the list of syntactically plausible candidates:

Domain Pertinence: let Di be the domain of interest (represented by a set of relevant documents) and let D_1,D_2, D_{i-1},D_{i+1},..D_N be sets of documents (or terminologies) in other domains, e.g. medicine, economy, politics, etc. The Domain Relevance of a term t wrt a domain Di is measured as[2]:

(1)
$$DR_{D_i}(t) = \frac{\hat{P}(t/D_i)}{\max\limits_{j}\left(\hat{P}(t/Dj)\right)} = \frac{freq(t,D_i)}{\max\limits_{j}\left(freq(t,Dj)\right)}$$

The Domain Pertinence is high if a term is frequent in the domain of interest and much less frequent in the other domains used for contrast. A similar filtering criterion is used also in (Park et al. 2002).

Domain Consensus: this measure, which is a novel measure that we introduced in (Navigli and Velardi, 2002), simulates the *consensus* that a term must gain in a community before being considered a relevant domain term. The Domain Consensus is an entropy-related measure, computed as:

[2] $\hat{P}$ is the Expected Value of the probability P.

(2) $$DR_{D_i}(t) = -\sum_{d_k \in D_i} \hat{P}(t/d_k)\log(\hat{P}(t/d_k)) = -\sum_{d_k \in D_i} norm\_freq(t,d_k)\log(norm\_freq(t,d_k))$$

where d_k is a document in D_i and *norm-freq* is a normalized term frequency. The domain consensus is then normalized for each term in the [0,1] interval. The consensus is high if a term has an even probability distribution across the documents chosen to represent the domain.

Lexical Cohesion: this measure evaluates the degree of cohesion among the words that compose a terminological string t. This measure is described in (Park et al. 2002) and proved to be more effective than other measures of cohesion in literature. Let $|t| = n$ be the length of t in number of words. The lexical Cohesion is measured as:

(3) $$LC_{D_i}(t) = \frac{n \cdot freq(t,D_i) \cdot \log(freq(t,D_i))}{\sum_{w_j} freq(w_j,D_i)}$$

where w_j are the words composing the term t. The cohesion is high if the words composing the term are more frequently found within the term than alone in texts.

Structural Relevance: if a term is highlighted in a document, e.g. it appears in the title or paragraph title, or if it is in bold or underlined etc., then the measure of its frequency, used in formulas (1-3) is increased by a factor k (user-adjustable).

Miscellaneous: a set of heuristics are used to remove from terms generic modifiers (e.g. *large* knowledge base), to detect mispellings[3] (using the WordNet on-line dictionary), to distinguish terminology from proper nouns, to extract single-word terminology, to detect acronyms, etc. We omit these details for sake of space, but the reader can access the web application and related documentation on http://lcl.di.uniroma1.it/termextractor.

The final weight of term is a linear combination of the three main filters:

(4) $$w(t,D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC$$

where the coefficients are user-adjustable, but the default is: $\alpha = \beta = \gamma = \frac{1}{3}$.

3. The web application

The TermExtraction web application has a pipeline architecture composed of 6 main phases, shown in Figure 2:

1. Set Termextractor options: in this phase the user can set several options or accept the default;
2. Upload documents: the user can upload documents in almost any format, one by one or in zipped archives;
3. Convert documents: documents are converted in txt format, heuristics are used to correct conversion errors (especially originated by pdf files);
4. Term Extraction: in this phase the terminology is extracted and filtered (see previous section);
5. Terminology Validation: in this phase a partner or a team of partners validate the terminology;
6. Save-download Terminology: in this phase the terminology is saved or downloaded in txt, xml or xcl format.

[3] Errors are often generated when converting documents in pdf or ps format in txt, see next Section.

At the end of phase 2 the user is disconnected[4], to allow for intensive data processing and to handle multiple users. At the end of extraction process (phases 3 and 4) the user receives an e-mail and is addressed to the validation page (phase 5). After the validation, the user can download the terminology, or he can save it on the TermExtractor server for further extension or validation (phase 6). The terminology is stored on the server for a limited time period (two weeks). We now provide some detail on each phase.

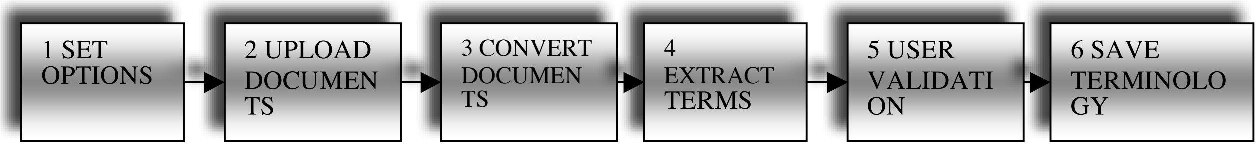


Figure 2. TermExtractor Pipeline Architecture

In **phase 1** the user is asked to set several options, or to accept the default options. For sake of brevity, we mention here only the most relevant settings[5]:

Select-deselect contrastive corpora: contrastive corpora are used to compute the Domain Relevance, through the formula (1). Examples of domains used for contrastive analysis are *medicine, computer networks*, etc. However, if the user domain of interest is, e.g. *wireless networks*, he may want to capture some more generic term in the area of computer networks. In this case, he can access the "Terminology" option and exclude *Computer Networks* from the list of domains used to compute the denominator of formula (1).

Set minimum and maximum length of terms: with this option the user can set the minimum and maximum length of multi-word terms to be extracted. Single-word terms are mostly (but not exclusively) extracted a posteriori, selecting the most frequent singleton components of multi-word terms.

Adjust the coefficients of the weight formula: here the user can tune the formula (4) by adjusting the three coefficients.

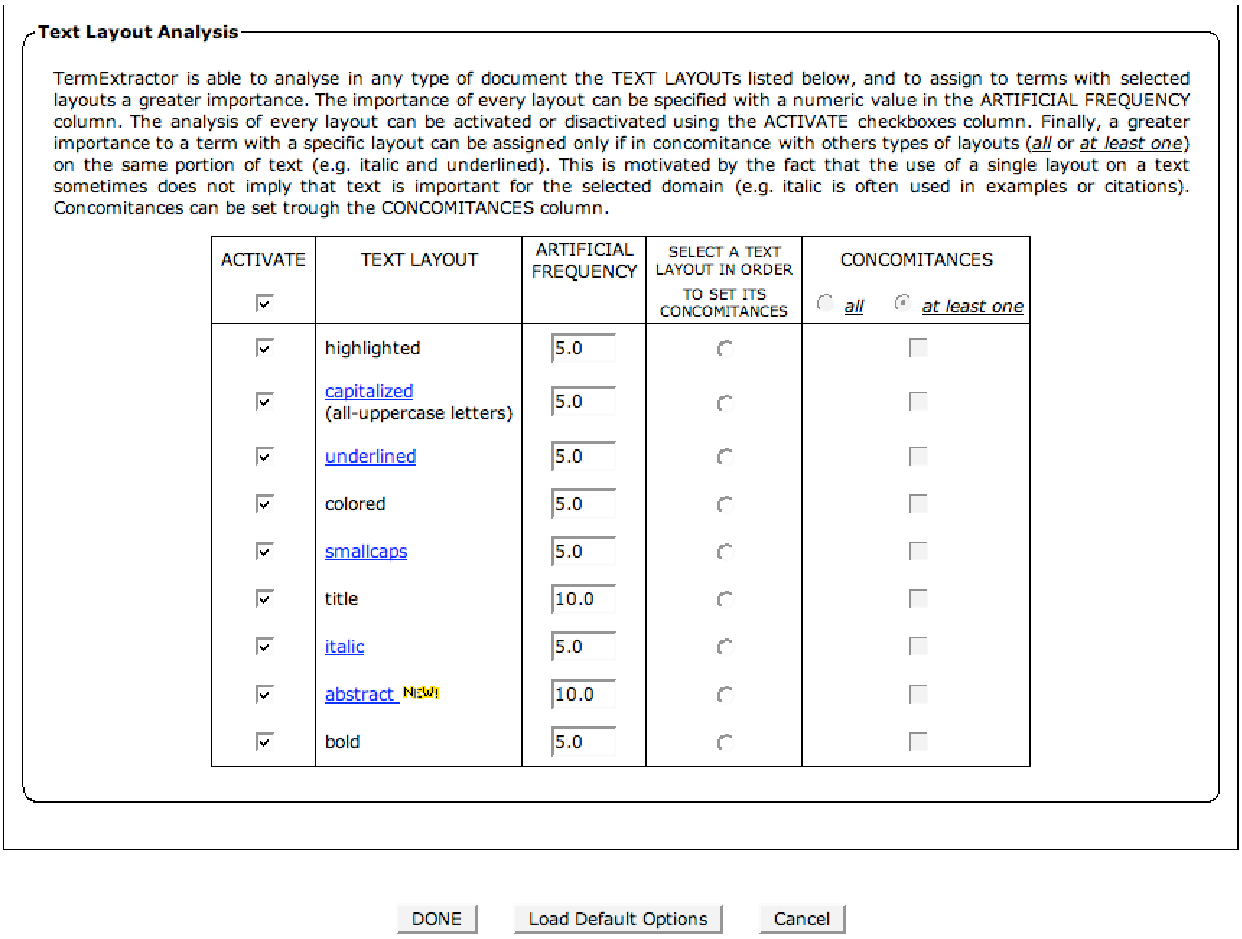
Other available options include: whether or not detecting proper nouns, whether to acquire a brand-new terminology, or enrich (and upload) an already existing one, selection of layout features to be considered in Structural Relevance analysis, etc.

Figure 3 shows one of the option windows, relative to layout analysis.

In **phase 2** the user can upload a set of documents that he considers relevant to model the domain under analysis. The effectiveness of TermExtractor filters depends on statistical significance, therefore in general, larger corpora obtain better results. Figure 4 shows the document-uploading interface. The user can upload up to 20 different documents, or as many documents he wants, compressed in a zipped archive. All main document formats are processed, as listed in the central box in Figure 4. It is also possible to specify the *url* of a web page. Once the documents have been uploaded, the user is disconnected from the application. When document processing is completed (**phases 3** and **4** of Figure 2), he receives an e-mail pointing to the web page where the evaluation can take place (**phase 5**). Single users can perform the evaluation, however, in real settings, a domain terminology must be evaluated and accepted *consensually* by a team of domain specialists. This is, for example, the evaluation procedure followed in the INTEROP EC project, as detailed in section 4.

[4] a "demo version" is available in which the user can upload a single document and obtain the result immediately.

[5] the interested reader can inspect the application and the available options on http://lcl.di.uniroma1.it

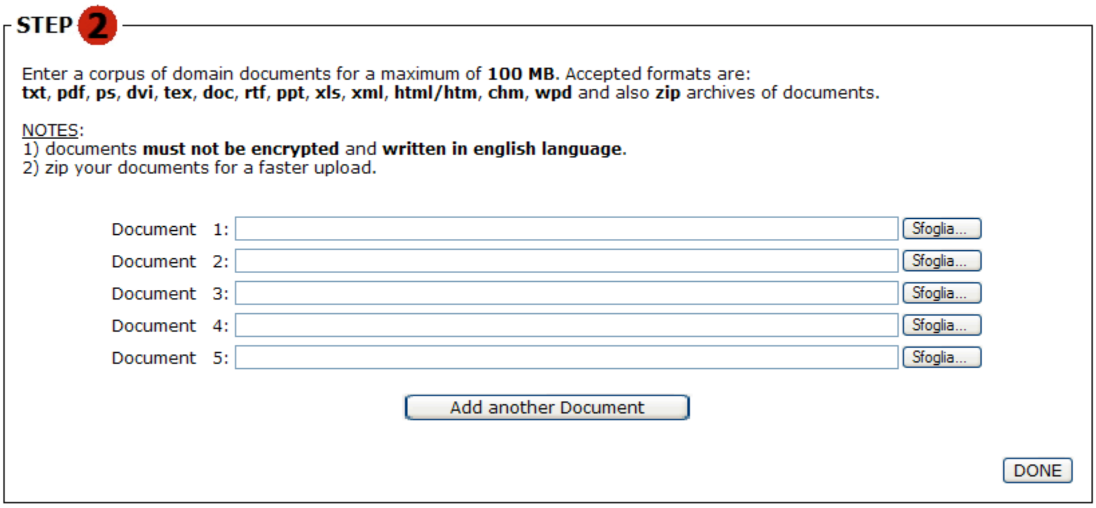


Text Layout Analysis

TermExtractor is able to analyse in any type of document the TEXT LAYOUTs listed below, and to assign to terms with selected layouts a greater importance. The importance of every layout can be specified with a numeric value in the ARTIFICIAL FREQUENCY column. The analysis of every layout can be activated or disactivated using the ACTIVATE checkboxes column. Finally, a greater importance to a term with a specific layout can be assigned only if in concomitance with others types of layouts (*all* or *at least one*) on the same portion of text (e.g. italic and underlined). This is motivated by the fact that the use of a single layout on a text sometimes does not imply that text is important for the selected domain (e.g. italic is often used in examples or citations). Concomitances can be set trough the CONCOMITANCES column.

ACTIVATE ☑	TEXT LAYOUT	ARTIFICIAL FREQUENCY	SELECT A TEXT LAYOUT IN ORDER TO SET ITS CONCOMITANCES	CONCOMITANCES ○ *all* ● *at least one*
☑	highlighted	5.0	○	□
☑	capitalized (all-uppercase letters)	5.0	○	□
☑	underlined	5.0	○	□
☑	colored	5.0	○	□
☑	smallcaps	5.0	○	□
☑	title	10.0	○	□
☑	italic	5.0	○	□
☑	abstract NEW!	10.0	○	□
☑	bold	5.0	○	□

Figure 3 . Setting options for layout analysis



STEP 2

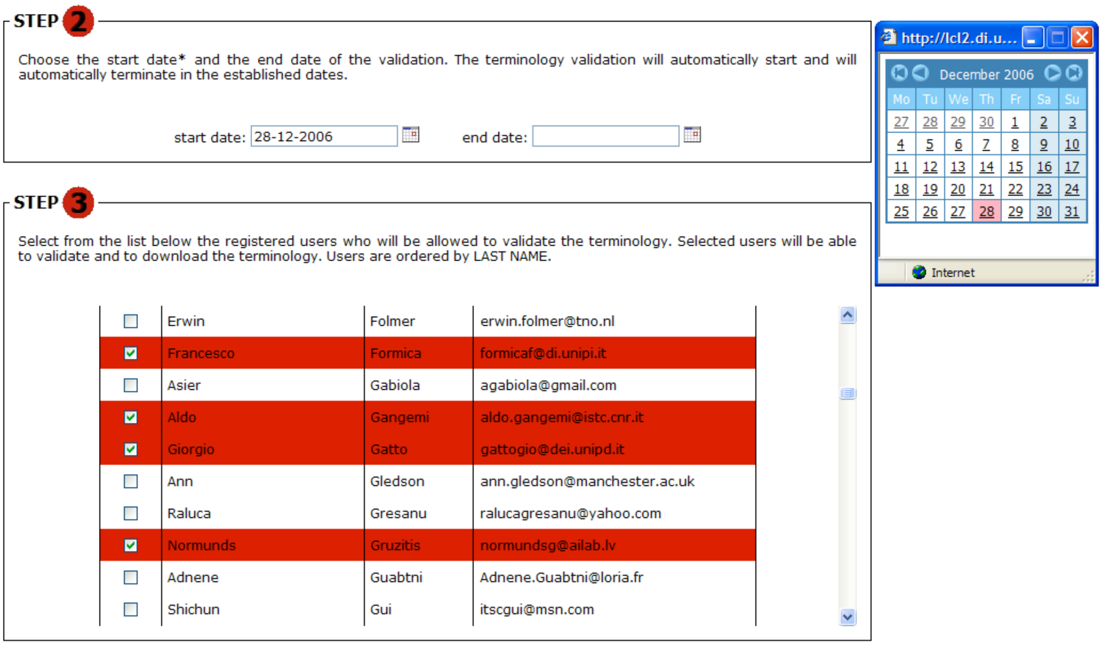
Enter a corpus of domain documents for a maximum of **100 MB**. Accepted formats are:
txt, **pdf**, **ps**, **dvi**, **tex**, **doc**, **rtf**, **ppt**, **xls**, **xml**, **html/htm**, **chm**, **wpd** and also **zip** archives of documents.

NOTES:
1) documents **must not be encrypted** and **written in english language**.
2) zip your documents for a faster upload.

Document 1:
Document 2:
Document 3:
Document 4:
Document 5:

Figure 4. Document uploading

Consensus building, which is particularly relevant when the objective is to learn the “domain lexicon” of a distributed community, is supported by a dedicated validation interface. A coordinator of the validation process must qualify himself, select the validation team, and establish the start and end date of the validation, as shown in Figure 5.



STEP 2

Choose the start date* and the end date of the validation. The terminology validation will automatically start and will automatically terminate in the established dates.

start date: 28-12-2006 end date:

STEP 3

Select from the list below the registered users who will be allowed to validate the terminology. Selected users will be able to validate and to download the terminology. Users are ordered by LAST NAME.

□	Erwin	Folmer	erwin.folmer@tno.nl
☑	Francesco	Formica	formicaf@di.unipi.it
□	Asier	Gabiola	agabiola@gmail.com
☑	Aldo	Gangemi	aldo.gangemi@istc.cnr.it
☑	Giorgio	Gatto	gattogio@dei.unipd.it
□	Ann	Gledson	ann.gledson@manchester.ac.uk
□	Raluca	Gresanu	ralucagresanu@yahoo.com
☑	Normunds	Gruzitis	normundsg@ailab.lv
□	Adnene	Guabtni	Adnene.Guabtni@loria.fr
□	Shichun	Gui	itscgui@msn.com

Mo	Tu	We	Th	Fr	Sa	Su
27	28	29	30	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Figure 5. Launching a team validation

Then, each partner of the validation team can inspect the terminology and accept or reject terms. The interface allows ordering extracted terms according to the global weight (formula 4), or to each of four indicators: Domain Relevance, Consensus, Cohesion and Frequency. Evaluators are also allowed to propose new terms and insert them in the list. This is shown in Figure 6. At the end of the validation period (but also during the validation), the coordinator can inspect the final result, as shown in Figure 7. The vote cumulated by each term is shown in the rightmost column.

In the final **phase 6**, the user can download the validated terminology in one of several formats (txt, xml, xls etc.). He can also temporarily save the terminology on the TermExtractor server, for further extensions.

4. Evaluation

One of the novel aspects of the work presented in this paper is evaluation. Unlike many existing systems for terminology extraction (see bibliography), TermExtractor has been validated in the large by web communities and individual users on different domains.

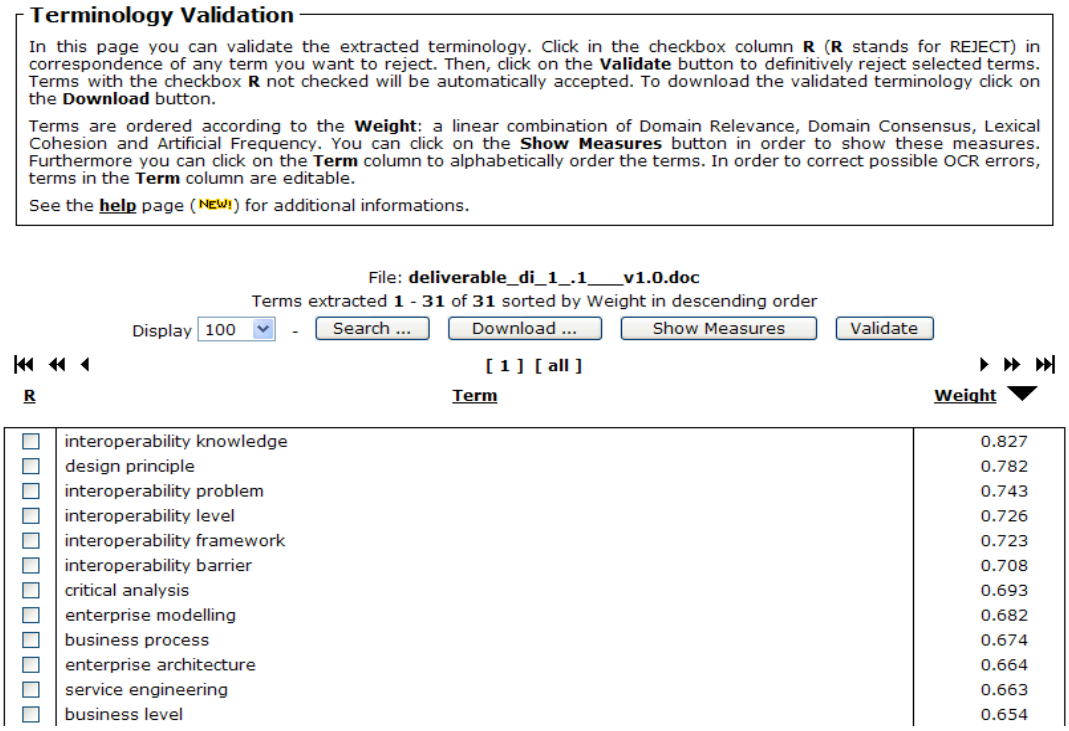
As previously mentioned, the term extraction algorithm (described in section 2) was first used within the EC INTEROP project[6] to extract a domain lexicon in the area of *enterprise interoperability research*. The *interoperability lexicon* was validated by the members of the Network of Excellence[7], and the results of the evaluation are shown in Table I. The extraction of the interoperability lexicon was the first step of the OntoLearn knowledge acquisition methodology, described in (Velardi et al. 2007), that eventually brought to the creation of a domain taxonomy[8]. Table I shows that the activity of term validation was greatly participated

[6] INTEROP is a network of excellence on enterprise interoperability research. The web site is http://www.interop-noe.org

[7] In INTEROP, the term extractor program has been used off-line. Only team validation was supported by a web application.

[8] The INTEROP taxonomy is browsable in http://lcl.di.uniromai.it/tav and http://interop-noe.org/backoffice/km/domains

by the NoE members, with around 2500 expressed votes. Overall, the precision of the system[9] was around 60%, but perhaps the most relevant result has been the speed-up of the lexicon creation process supported by TermExtractor. In emergent web communities, like the INTEROP NoE, often the domain of interest is not well-assessed, therefore the task of manually identifying the relevant domain terms by a team of specialists is difficult and time-consuming. In INTEROP, the community was composed by researchers belonging to rather heterogeneous fields, namely: *ontology* and *knowledge representation*, *enterprise modeling*, *architectures* and *platforms*. The "enterprise interoperability" domain was initially weakly defined, therefore capturing the common, relevant concepts was one of the first targets of the NoE. TermExtractor fostered the identification of many emergent domain concepts, and furthermore it provided a valuable support to consensus building, which is a key issue during the definition of a domain terminology.



Terminology Validation

In this page you can validate the extracted terminology. Click in the checkbox column **R** (**R** stands for REJECT) in correspondence of any term you want to reject. Then, click on the **Validate** button to definitively reject selected terms. Terms with the checkbox **R** not checked will be automatically accepted. To download the validated terminology click on the **Download** button.

Terms are ordered according to the **Weight**: a linear combination of Domain Relevance, Domain Consensus, Lexical Cohesion and Artificial Frequency. You can click on the **Show Measures** button in order to show these measures. Furthermore you can click on the **Term** column to alphabetically order the terms. In order to correct possible OCR errors, terms in the **Term** column are editable.

See the **help** page (NEW!) for additional informations.

File: **deliverable_di_1_.1___v1.0.doc**

Terms extracted **1** - **31** of **31** sorted by Weight in descending order

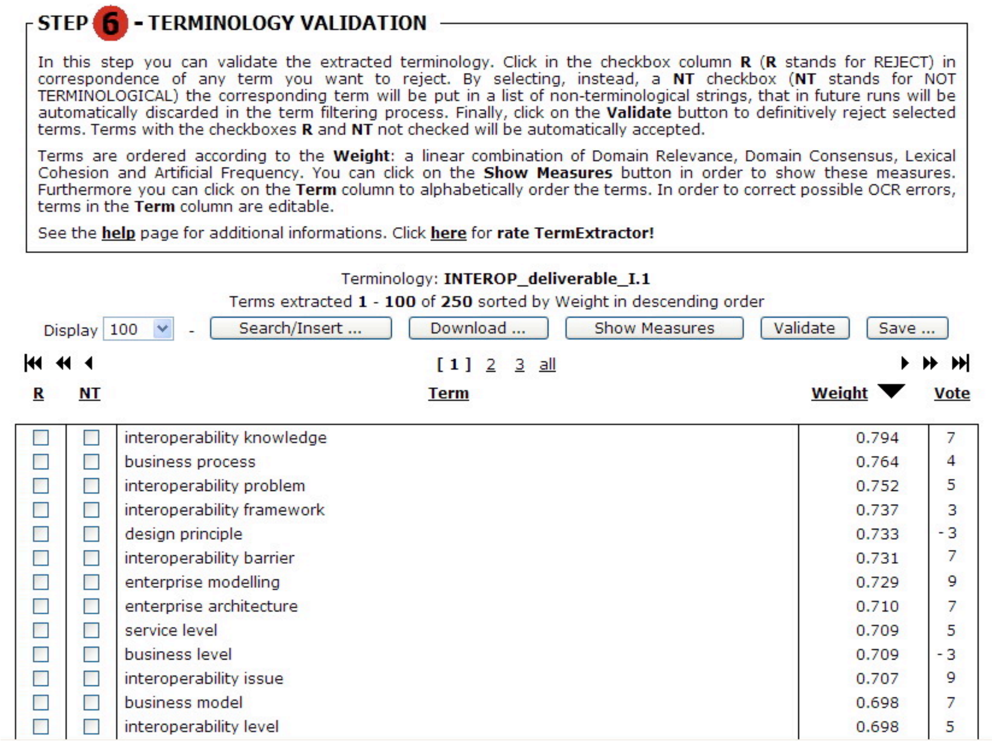
R	Term	Weight
□	interoperability knowledge	0.827
□	design principle	0.782
□	interoperability problem	0.743
□	interoperability level	0.726
□	interoperability framework	0.723
□	interoperability barrier	0.708
□	critical analysis	0.693
□	enterprise modelling	0.682
□	business process	0.674
□	enterprise architecture	0.664
□	service engineering	0.663
□	business level	0.654

Figure 6. Validation by a single partner

After the validation of the term extraction methodology in INTEROP, it was decided to create a web application, to make the term extraction service available outside the Network of Excellence, and to demonstrate the generality of the approach. The results of the INTEROP collective validation and suggestions by users were exploited to add further improvements to the system, which was uploaded and made freely available in October 2006 on http://llcl.di.uniroma1.it.

To evaluate the precision and the user-friendliness of the application, we asked several institutions around the world to test the system on domains at their choice, and provide a global judgment.

[9] note that the evaluation of TermExtractor from the INTEROP NoE took place on November 2005. After that date, many improvements in the algorithm have been added.



STEP 6 - TERMINOLOGY VALIDATION

In this step you can validate the extracted terminology. Click in the checkbox column **R** (**R** stands for REJECT) in correspondence of any term you want to reject. By selecting, instead, a **NT** checkbox (**NT** stands for NOT TERMINOLOGICAL) the corresponding term will be put in a list of non-terminological strings, that in future runs will be automatically discarded in the term filtering process. Finally, click on the **Validate** button to definitively reject selected terms. Terms with the checkboxes **R** and **NT** not checked will be automatically accepted.

Terms are ordered according to the **Weight**: a linear combination of Domain Relevance, Domain Consensus, Lexical Cohesion and Artificial Frequency. You can click on the **Show Measures** button in order to show these measures. Furthermore you can click on the **Term** column to alphabetically order the terms. In order to correct possible OCR errors, terms in the **Term** column are editable.

See the **help** page for additional informations. Click **here** for **rate TermExtractor!**

Terminology: **INTEROP_deliverable_I.1**

Terms extracted **1** - **100** of **250** sorted by Weight in descending order

Display 100

R	NT	Term	Weight	Vote
□	□	interoperability knowledge	0.794	7
□	□	business process	0.764	4
□	□	interoperability problem	0.752	5
□	□	interoperability framework	0.737	3
□	□	design principle	0.733	- 3
□	□	interoperability barrier	0.731	7
□	□	enterprise modelling	0.729	9
□	□	enterprise architecture	0.710	7
□	□	service level	0.709	5
□	□	business level	0.709	- 3
□	□	interoperability issue	0.707	9
□	□	business model	0.698	7
□	□	interoperability level	0.698	5

Figure 7. Inspecting the results of a team validation

n. of extracted terms	1902
Total voting partners	35
Total expressed votes	2453
Total different terms with a negative vote	783 (41%)
Survived terms	1120

Table 1. Summary of INTEROP collaborative terminology validation (November 2005)

The team of evaluators includes a restricted group of INTEROP partners and was enriched with web users showing some interest in the application and volunteering to participate in the evaluation. The results shown in Table II are rather encouraging. In Table II, INTEROP members are highlighted (but the documents used for test where different from those used to extract the INTEROP lexicon), the other are external users who volunteered to perform the evaluation. Users 2 and 5 are private companies, the other are research institutions. In the table, the precision is automatically computed by the system, depending upon the number of rejected terms, while the global judgment is specified by each evaluator, selecting in a dedicated window from 5 possible judgments ranging from very good to bad.

5. Related research

Most available on-line term extraction services[10] use very simple extraction algorithms: typically, they extract every word and every phrase up to a certain number of words in length that occurs at least a minimum number of times in a source text file and that do not start or end with a stop word. The document to be analyzed can either be a web page, specified trough its *url*, or a plain text submitted by the user.

[10] e.g. Topicalizer http://www.topicalizer.com/ and Textalyser http://textalyser.net/

To the best of our knowledge, the only terminology learning application with comparable complexity of the extraction algorithms is the IBM Glossex system (Park et al. 2002), (Kozakov et al. 2004). Similarly to TermExtractor, Glossex filters terminological candidates using lexical cohesion and a measure of domain relevance. Glossex has also some additional useful heuristics, like aggregation of lexical variants for a term (e.g. compunding variants like *fog lamps* and *foglamps*, or inflectional variants like *rewinding* and *rewound*). On the other side, Glossex analyses one document at a time, therefore it is unable to identify popular domain terms with an even probability distribution across the documents of a collection (like we do with Domain Consensus measure). Unfortunately, a performance comparison based on the analysis of single documents is not possible since the Glossex tool is not freely available.

We finally stress that in virtually all papers on terminology extraction mentioned so far the validation is conducted manually by three judges (usually the authors themselves). This is not comparable with the large-scale evaluation of TermExtractor conducted within web communities like INTEROP, and by several external institutions around the world, on different domains and competences.

References

BOURIGAULT D. AND JACQUEMIN C.: Term Extraction+Term Clustering: an integrated platform for computer-aided terminology, in Proc. of EACL , 1999

COLLIER N., NOBATA C. AND TSUJII J. : Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain, Terminology, 7(2). 239-257, 2002

KOZAKOV,L. Y. PARK, T. FIN, Y. DRISSI, Y. DOGANATA, AND T. COFINO "Glossary extraction and utilization in the information search and delivery system for IBM Technical Support", IBM System Journal, Volume 43, Number 3, 2004

R. NAVIGLI, P. VELARDI. Semantic Interpretation of Terminological Strings, Proc. of *6th International Conference on Terminology and Knowledge Engineering* (TKE 2002), INIST-CNRS, Vandoeuvre-lès-Nancy, France, August 2002, pp. 95-100

NAVIGLI R. AND VELARDI, P. (2004). "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites". Computational Linguistics. vol. 50 (2).

NAVIGLI R., P. VELARDI "Structural Semantic Interconnections: a knowledge-based approach to word sense disambiguation" Special Issue-Syntactic and Structural Pattern Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 27, n. 7, 2005

PARK Y. , R. J. BYRD, B. BOGURAEV "Automatic glossary extraction: beyond terminology identification" International Conference On Computational Linguistics , Proceedings of the 19th international conference on Computational linguistics – Taipei, Taiwan, 2002

P. VELARDI A. CUCCHIARELLI AND M. PÈTIT "A Taxonomy Learning Method and its Application to Characterize a Scientific Web Community" IEEE Transaction on Data and Knowledge Engineering (TDKE), vol. 19, n. 2, February 2007, pp. 180-191

WERMTER J. AND HAHN U.: Finding New terminology in Very large Corpora, in Proc. of K-CAP'05, October 2-5, 2005, Banff, Alberta, Canada

Company/ Organization[11]	Vote[12]	Terminology description (as Indicated by user)	# of documents	# of terms before validation	#r of terms after validation	Precision[13]
The Pavel Terminology Tutorial - Translation Bureau - Government of Canada	Very Good	A chapter from a Collective Agreement.	1	39	37	0.948
Stockholm University	Very Good	English anatomy/medical terms.	7680	4457	4432	0.994
K-Tech	Very Good	Terminology about design pattern from 5 different books	5	549	534	0.973
IASI - Institute of Systems Analysis and Computer Science. Rome, Italy. (INTEROP partner)	Good	Ontology alignment, Ontology mapping, Ontology matching, model transformation	58	157	145	0.923
School of Information Technology and Engineering. University of Ottawa, Canada.	Very Good	Terminology extracted from the Nuclear-Free Canada program.	7	575	513	0.892
Computational Linguistics Research. Damascus, MD, US.	Very Good	Terminology developed from Senseval-3 papers.	19	58	49	0.844
Università Politecnica delle Marche, Ancona, Italy. (INTEROP partner)	Good	Security Protocols	40	158	128	0.810
Jena University Language and Information Engineering (JULIE) Lab, Germany	Good	Immunology (It's a broad branch of biomedical science that covers the study of all aspects of the immune system in all organisms).	1	35	25	0.714
Federal University of Ceará	Very Good	Embodied cognitive science	5	131	85	0.648
Università Politecnica delle Marche, Ancona, Italy. (INTEROP partner)	Good	Semantic Similarity	16	80	49	0.612
CIMOSA	Good	Standardisation ISO TC 184 SC5 and related activities	18	102	62	0.607
Universidad Politecnica de Valencia, Spain. (INTEROP partner)	Good	Collaborative Networks from the point of view of Enterprise Modelling	63	115	60	0.521
Waikato University, New Zealand.	Good	Semantic Web and Ontologies	1	42	35	0.833
Department for Computational Linguistics, University of Potsdam.	Very Good	"Thinking in C++" Vol.1 - Bruce Eckels	32	237	231	0.974
						Average Precision 0.806

Table II- Evaluation of the TermExtractor web application by several web users

[11] Volunteers accepted to publish the name of their institution

[12] Possible votes were: Very Good, Good, Fair, Poor and Bad.

[13] Precision = (Number of terms before validation) / (Number of terms after validation)