Multilingual Access to Library Catalogues: Word Sense Disambiguation via Classification Systems

Raffaella Bernardi[1], Daniele Gobbetti[1], and Luigi Siciliano[2] *

[1] Free University of Bozen-Bolzano, Faculty of Computer Science – KRDB
bernardi,gobbetti@inf.unibz.it,
[2] Free University of Bozen-Bolzano, University Library
luigi.siciliano@unibz.it

Abstract. Multilingual Digital Libraries ask for Cross-Language search engines able to disambiguate the search terms across languages. In this paper, we report on the development and first evaluation of a Cross-Language Information Retrieval (CLIR) system enriched with a Word To Category Module: we automatically extract a mapping from words in the meta data of the Catalog's records to the associated Classification Categories and exploit it to help the CLIR system retrieve only books in different languages about the topic actually queried by the user.

1 Introduction

Polysemic words challenge the access to multilingual catalogues and the application of machine translation to such task. If an end-user searches for records by typing the Italian keyword "banca" (ie. an establishment for the custody, loan, exchange, or issue of money), a cross-language information retrieval (CLIR) search engine will query the catalogues by using also the English translation of the Italian keyword, namely "bank". The latter, however, is a polysemic word and has meanings that do not belong to the source language word "banca". Some sort of Word Sense Disambiguation (WSD) needs to be applied so to retrieve only English books about the topic associated to the source language, viz. Economics.

Resnik and Yarowsky [RY00] conducted an in-depth empirical study of translingually-based annotations of polysemous contexts across 12 languages. They show that the probability that a given sense pair will tend to lexicalize differently across languages is correlated with semantic salience and sense granularity. Their results are summarized in the Table 1 where it is given the mean probability that a given language differently lexicalizes an English sense distinction in the Hector inventory [Atk93]. Both Indo-European (IE) and non-Indo-European (NI) languages are considered. As the authors remark one implication of these results for machine translation is that for more distant languages from English, such as Japanese, around 86% of the monolingual sense distinctions correspond to translation distinction and hence need to be resolved in Machine Translation (MT). For relatively similar languages, such as Spanish-English, the importance of WSD is apparently lower, given that approximately 50% of the sense distinctions noted by lexicographers need not be resolved due to parallel polysemy in the target language. Nevertheless, also this value is high enough to warrant some form of WSD for lexical choice in MT systems.

In this paper, we present our work in progress on the development of an automatic mapping between words occurring in the meta-data of a Multilingual Library Catalogue and Classification Numbers (of the Classification System used to catalogue the Library records), e.g. the Dewey Decimal Classification (DDC[3]). The mapping will serve as a support module to a CLIR system to

* This work is supported by the eContent*plus* ECP 2006 DILI 510035. We thank Emanuele Pianta and Yannick Versley for providing us with a list of polysemic words in the three languages of our study; Marco Baroni and Massimo Poesio for pointing us to relevant literature on WSD. Furthermore, we would like to thank all the people who helped us with the experiments discussed in this paper. We acknowledge the FBK team for letting us use the WordNet Domain with a Research Licensee. Finally, we warmly thank the Universitätsbibliothek Regensburg for having provided an XML version of the Regensburger Verbundklassifikation (RVK) Classification System.

[3] http://www.oclc.org/dewey/

Language	Avg P$_L$	Language	Avg P_L
NI - Basque	0.885	IE - Romanian	0.667
NI - Japanese	0.856	IE - Greek	0.635
NI - Korean	0.846	IE - Hindi	0.558
NI - Chinese	0.808	NI - Arabic	0.538
NI - Turkish	0.692	IE - Spanish	0.500
NI - Hungarian	0.692	IE - Swedish	0.461

Table 1. Different lexicalization of English polysemic words

help retrieving only those books about the relevant topic, i.e. those books that have been assigned classification numbers shared by the words of the source and target languages. For instance, the Italian word “banca” could be associated to DDC 332.1 (Social science, Economics, Financial economics, Banks), whereas its English translation “bank” both to DDC 332.1 and DDC 627.133 (Technology, Engineering, Hydraulic engineering, Inland waterways, Canals, Bank protection and reinforcement). Only English books associated with the shared classification number (or relatives of it) should be returned. In Section 2, we present a brief overview of the related literature. In Section 3, we describe the Architecture of a CLIR system and the role of our module in it. Its evaluation is discussed in Section 4 which is the base for the conclusion drawn in Section 5.

2 Related Works

WSD has attracted the attention of researchers in computational linguistics since the earliest days of computer treatment of language in the 1950’s, and nowadays campaigns, like SensEval, have been set up to evaluate WSD systems and highlight their strengths and weaknesses. In these settings, WSD is seen as the task of assigning the most appropriate meaning (sense) to a polysemous word within a given context. Similarly, the problem of WSD within MT is tackled as the task of detecting the word in the target language and the word is disambiguated by the context it occurs in. In our scenario, instead, users enter one or at most two keywords in a source language that set the sense of interest for disambiguating the polysemic word of the target language. Hence, we cannot exploit co-occurring words to detect the relevant sense. On the other hand, our task is simplified by the association of records with classification categories that are used trans-nationally and hence can be exploited as an intermediator between the target and source languages. *Classification categories* could be seen as indication of sense distinctions.

The idea of using a mapping of words to categories is discussed already by Yarowsky in [Yar92]. The author proposes an approach to WSD that uses classes of words to derive models useful for disambiguating individual words in context. To this end, he takes Roget’s categories as an approximation of conceptual classes and hence as sense distinctions. The author shows that selecting the most likely category provides a useful level of sense disambiguation. For our task, though, the classes of words must be based on topic relations.

The idea of using macro categories for WSD is further exploited in [MSPG98]. The work is based on the “one domain per discourse” hypothesis. In other words, the underlying idea is that domain labels, such us “Medicine”, “Architecture” and “Sport”, provide a useful way to establish semantic relations among word senses, which can be profitably used during the disambiguation process. A domain is a set of words between which there are strong semantic relations. Like in [MSPG98], we are interested in capturing information about the *domain* to which a word belongs to, in order to disambiguate it. Again, differently from their task, in our the word to disambiguate is just a keyword without context. Yet, we will take advantage of the lexical resource developed by the authors, viz. “WordNet Domain” (WND).

WND is an extension of WordNet[4] in which synsets have been annotated with one or more domain labels, selected from a hierarchically organized set of about two hundred labels. WordNet

[4] http://wordnet.princeton.edu/

Domain has been built starting from about 200 domain labels selected from a number of dictionaries and then structured in a taxonomy according to the DDC. The resulting taxonomy is a sub-graph of DDC, e.g. the words in {beak, bill, neb, nib} have been annotated with "Zoology" starting from the synset "bird" and following a part-of relation. In the case of WordNet synsets that do not belong to specific domain, but rather appear in tests associated with any domain, they have been labeled as "Factotum" (it includes generic sysnsets that are hard to classify in a particular domain, such as "man", and stop sense synsets that appear frequently in different contexts, such as "numbers", "colors".) Similarly, we are interested in clustering words by *topic*; we consider the membership of a word to a domain to be a relevant information for detecting its membership to a *topic*.

Relevant work on topic disambiguation has been conducted in the Digital Library field. In [GS04], Giffiths and Steyvers propose a statistically based *topic model* to identify the content of a document. Our aim is similar, but we will follow a more pragmatic approach and see how a CLIR search engine could be improved in its multilingual search over a Multilingual Library Catalogue by simply clustering content words (verbs, nouns and adjectives) in the meta data of books with the same classification number and use the obtained mapping between the cluster of words and the classification number for topic disambiguation and hence help the multilingual search engine.

3 The W2C module within a CLIR system

Our work will be part of a CLIR system, to be delivered by an EU project. In this section, we describe the system architecture, describe the Word To Category (W2C) module and evaluate it. In Figure 1, we give an overview of the CLIR system architecture.

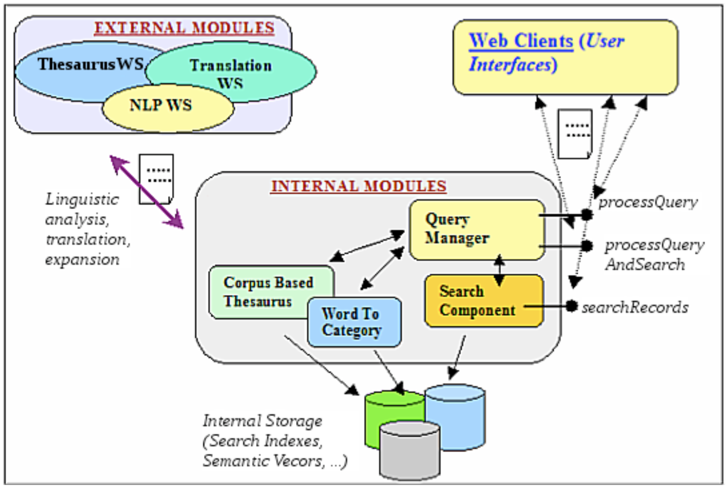

Fig. 1. CLIR system

Given a record we extract a mapping between the words occurring in its metadata and the classification number associated to it. By repeating this extraction for all records in the Catalogue we build a mapping between clusters of words and classification numbers. We expect the classification number to be a proper indicator of the sense of the words in the cluster associated to it, hence the mapping can be exploited to identify the proper sense of a target word and filter out the others In this paper we report our evaluation on the W2C by applying it to the Library Catalogue of our University.

The Library holds 139,481 titles classified with the "Regensburger Verbundklassifikation" (RVK[5]) (nr of records per top RVK categories: 25,599 in Q (Economy), 23,008 in D (Pedagogy), 13,160 C (Philosophy and Psychology), 10,653 in M (Politics and Sociology), 6,100 in ZG-ZS (Technics. Architecture. Engineering), all the other categories have less than 5,000 each). Books

[5] http://www.bibliothek.uni-regensburg.de/rvko_neu/

are labeled with different Subject Headings on the base of their language: The German records of the Library are labelled with Subject Headings (SH) of the "Schlagwortnormdatei" (SWD) vocabulary, the Italian and English books are labelled with SH of the "Soggettario Italiano" and the "Library of Congress Subject Headings", respectively. RVK hierarchy contains 33 top categories which can be easily mapped into the domains of WND[6], we will use this mapping in our evaluation.

Titles have been tagged with part-of-speech. All words have been extracted from the titles and subject headings and associated with the classification number of the book in whose metadata they occur. The clusters have been cleaned by leaving in only verbs, adjectives and nouns and filtering out other PoS classes. Words are listed in the cluster by absolute frequency, see the example in Table 2. (We plan to consider other measures, like mutual information, in the future.) The words within a same cluster are expected to be semantically related – they are not necessarily taxonomicaly similarly but might be linked by a broad set of relations, such as meronomy, functional, object of an action, modifier of a noun – and anyway are expected to belong to the topic of the classifier number associated with it.

words from title	RVK	WND domain
field, bonds, commitments, venture, giving, affiliations, gross, incorporating, trust, club, valuation, stagnation, office, selling, manufacturer, saving, move, dollar, balance, airline, preferred, infrastructures, foundation, industrialism, deal, operation, land, fraud, issue, provision, account, pockets, organization, agencies, got, taxation, getting, exchange, bankruptcies, reorganization, accession, traded, promotion, administrations, reward, takes, assessing, transactions, cost, capacity, ...	Q	Economics

Table 2. Cluster of words associated to the category Q, word domain "Economics"

In WordNet, the polysemic word "bank" is said to have the following sense 1. "sloping land", 2. "a financial institution that accepts deposits"; 3. "a long ridge or pile"; 4. "an arrangement of similar objects in a row or in tiers"; 5. "a supply or stock held in reserve for future use", 6. "the funds held by a gambling house"; 7. "a slope in the turn"; 8. "a building in which the business of banking transacted"; 9. "a flight maneuver". The senses that have a clear "topic cue" are on the one hand 2. and 8. and on the other 1. that clearly belongs to "Economics" and "Earth", respectively. The other senses of the words could be used in different contexts, i.e. belong to several topics. Our W2C module associates "bank" to several RVK categories as illustrated in Table 3 where are also given all the domains to which the word belongs in WND (WND domain). From other examples we have scrutinized, it seems that the senses which have a strong membership to a specific topic/domain are those that are both in the RVK and in the domain of the WND[7]. We will investigate this further and, if the observation is confirmed, we will take advantage of it to fine tune the W2C module and increase its precision.

The 5451 clusters generated with words taken from english titles have a size of 18.39 (on average) words each. Half of the words (i.e. 9.20 on average) are part of the factotum domain of the WND. The clusters made with words taken from the Subject Headings are smaller in size (9.60 words on average) but none of the words are in the factotum domain.

4 Evaluation

To evaluate the system we proceeded into two steps. We evaluated (1) the quality of the clusters obtained; (2) the functionality of the extracted mapping for disambiguating the target words resulted from the translation. In both cases we conducted controlled tests over a sample of clusters and queries, respectively, by asking human beings evaluations. We asked to researchers and librarians specialized in different disciplines. For each language, at least one of the user was a native speaker.

[6] Only "Archeology" is not accounted for in the WND, and a few WND domains do not have a mapping into RVK.

[7] Notice, our Library does not have books on Geography (Earth).

RVK, mapped domain	**WND domain**	**WN sense**
H, Literature		
MA-ML, Politics		
P, Law		
Q, Economics	Economics	2. a financial institution that accepts deposits
		8. a building in which the business of banking transacted
SA-SP, Mathematics		
ZA-ZE, Home-Agriculture-Food		
	Earth	1. sloping land
	Sport	
	Transport	

Table 3. Bank's RVK and WordNet Domains

	Category, Label	Nr. of words	Correct tagging
Italian cluster	CX 5000, Psychology	190	0%
	MS 8020, Sociology	191	60%
	DT 1000, Pedagogy	114	60%
	YM 6100, Medicine	6	20%
	HD 120, English studies	11	0%
German cluster	DO 1260, Pedagogy	128	60%
	ZH 8900, Technology	67	20%
	MS 2850, Sociology	458	100%
	ZD 50000, Home, Agriculture..	12	80%
	WI 2500, Biology	15	20%
English cluster	QR 300, Economics	143	100%
	NC 2009, History	5	0%
	ST 237, Computer Science	45	100 %
	MD 4700, Politics	31	25%
	LH 79500, Art	220	0%

Table 4. (a) Cluster to Category Label

Quality of the clusters We are interested in understanding whether the classification number assigned to a cluster is a good indicator of the words' sense in the associated cluster, and hence, whether the words do belong to the same topic. Experiments (a)-(c) help us answering this question.

(a) Cluster to Category Label Given a cluster of words and the list of Category Labels the user (6 users per language) was asked to assign to the cluster a label from the list.[8] (See results in Table 4).

(b) Word to Category Label Given two unrelated RVK labels and a list of words (all belonging to the cluster associated with one of the two RVK label), for each word the user grades its relationship to the RVK labels (mark from 0-4).[9] (See results in Table 5).

(c) Word to Category Label Hierarchy Given two unrelated RVK labeled sub-hierarchies and a list of words (all belonging to the cluster associated with one of the two RVK label), the user has to assign each word to the RVK and indicate the exact sub-node too.[10] (See results in Table 6).

As we can see from Table 4, the result was worse when the cluster contained fewer words. In most of the cases of wrong tagging, the user has chosen category strongly related with the

[8] The clusters have been generated making sure that they contain some polysemic words with high frequency and that were not associated to the A-RVK category, viz. "General".

[9] The words were presented in alphabetic order and each list contained max. 40 words.

[10] This test has been carried out only for German, since it requires all the sub-categories label of the RVK system.

English words	ZB 60406	DS 6000	WK 9000	ZP 3760	CC 7200	ST 270
	Home, Agriculture . . .	Pedagogy	Biology	Technology	Philosophy	Computer Sci.
correct	32.69%	20.00%	42.86%	55.56%	54.21%	68.12%
incorrect	7.69 %	10.00 %	42.66%	41.67%	22.43%	13.09%
draw	59.62%	70.00%	14.29%	2.78%	18.69%	18.12%

Italian words	IV 51280	MS 6440	CQ 5100	QQ 925	QH 400	CI 5624
	Romance studies	Sociology	Psychology	Economics	Economics	Philosophy
correct	24.58%	59.51%	54.65%	53.16%	66.67%	42.86%
incorrect	23.46%	9.27%	16.28%	15.19%	33.33%	53.57%
draw	51.96%	31.22%	29.07%	31.65%	0.00%	3.57%

German words	MS 4710	ZH 1350	ML 1000	LP 65658	NQ 4300	WK 7300
	Sociology	Technology	Politics	Musicology	History	Biology
correct	61.76%	34.69%	91.30%	100.00%	100.00%	75.00 %
incorrect	11.76 %	22.45%	8.70%	0.00%	0.00%	25.00%
draw	25.00%	36.73%	0.00%	0.00%	0.00%	0.00%

Table 5. (b) Word to Category Label

German words	DX 4450	ZH 4800	DD 9231	MG 70070	UB 4085	MQ 2700
	Pedagogy	Technology	Pedagogy	Politics	Physics	Sociology
correct	9.64%	15.35%	53.85%	100.00%	100.00%	56.99%
incorrect	3.61%	2.33%	46.15%	0.00%	0.00%	5.38%
draw	10.84%	14.88%	0.00%	0.00%	0.00%	35.48%

Table 6. (c) Word to Category Label Hierarchy

right one, e.g., "Pedagogy" and "Psychology" have been interchanged, similarly, "Medicine" with "Biology" and "Economics" with "Social Science". In the latter case, the error would have not come across if we had used a classification system with fewer and more general topmost elements such as DDC, where for instance Economics and Social science are under the same class.

Quality of the translation The evaluation of the support to the translation system has been conducted by manually checking a sample of words (as well as the corresponding translation proposals) and their inclusion as candidate words before and after enabling the module. The behavior was marked as correct in 48% of the cases, the exclusion of candidates seems to be slightly poor since it was considered appropriate in 53% among the evaluated translations. The results might be due to the characteristics of the RVK that contains many macro-categories and unbalanced distribution among the topics; the small size of the catalogues we used as well as its focus on some topics; the fact that the words have been extracted only from titles without considering the subtitles and the subject headings.

5 Conclusion and Further Research

The evaluation we have carried out so far shows that the W2C is a promising module within a CLIR system. What needs to be studied more in depth is (a) the nature of words in the cluster, (b) the size of the clusters associated to the Classification number, and (c) the level of granularity of the Categories to be considered. We plan to answer these questions by conducting a study of the Library's logs; evaluate the precision and recall of the CLIR system enhanced with the W2C module over a fixed set of queries (e.g. those used at CLEF-TEL); verify the importance of keeping compound nouns in the cluster. Based on the results of these further experiments, we will consider using WND to reduce the size of the clusters and improve their quality and hence precision of the system. Finally, we will apply the module on other Libraries and test it over different Classification Systems too.

References

[Atk93] S. Atkins. Tools for computer-aided lexicography: the hector project. In *Papers in Computational Lexicography: COMPLEX'93*, 1993.

[GS04] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[MSPG98] B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 1(1), 1998.

[RY00] P. Resnik and D. Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3), 2000.

[Yar92] D. Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, 1992.