Extraction and mapping of CIDOC-CRM encodings from texts and other digital formats

M. Généreux[1] and F. Niccolucci[2]

[1]University of Brighton, United Kingdom
[2]PIN, Prato, Italy

Abstract

CIDOC-CRM is a new standard for encoding a wide range of information for Cultural Heritage (CH). At present, existing CH collections are stored using all sorts of formats, sometimes proprietary, often defined roughly, which makes it difficult to share or access heterogeneous information among the CH community. There is a need for a tool to map diverse formats into CIDOC-CRM, assisted by another tool using intelligent language technology to help the mapping whenever fields are underspecified or loosely described, both tools being complementary. In some cases, it may even be better to build fragments of a CIDOC database directly from informal descriptions in natural language only, as the CH community may be reluctant to switch to new formats of data entry. Therefore, this paper focus primarily on the mapping of CH data described in natural language into CIDOC-CRM triples, the building blocks of the full CIDOC-CRM ontology. The methods exploits the propositional nature of CIDOC-CRM triples. Using WordNet as a lexical database and the WEB as corpus, we first extract triples from examples provided in the CIDOC-CRM literature, and then from text describing the medieval city of Wolfenbüttel. We show the strong points of the system and suggest where and how it could be improved. Although the triples extracted automatically from texts do not provide a full picture of the CIDOC-CRM structure buried in the textual description, our results indicate that it provides a sound initial working basis for the mapping/translation process, saving time on what would otherwise have to be done by hand.

Categories and Subject Descriptors (according to ACM CCS): J.5 [Computer Applications]: Arts and Humanities

1. Introduction

Like it or not, the CH community will have to get acquainted to a new ontology for storing databases and collections. The CIDOC-CRM ontology aims at accommodating a wide variety of data from the CH domain, but it sheer complexity may make it difficult for non-expert to learn it quickly, let alone use it efficiently. For others, it may even be simpler to find a way to translate automatically their data from the storage mechanism already in place into CIDOC-CRM. For practitioners unfamiliar with tight formalisms, it may be more natural to describe collections in natural language (e.g. English), and there is already an unprecedented wealth of information available on-line in natural language for almost anything, including CH. Wouldn't it be practical to be able to describe a collection of artifacts in plain English, with little or no knowledge of the CIDOC-CRM formalism, and

use language technology take over and produce a CIDOC-CRM database? This paper presents a method to do just that. It is based on the idea that the building blocks of the CIDOC-CRM ontology, the *triples*, have a predicative nature, which is structurally consistent with the way many natural languages are build. According to [CIDb]:

> The domain class is analogous to the grammatical subject of the phrase for which the property is analogous to the verb. Property names in the CRM are designed to be semantically meaningful and grammatically correct when read from domain to range. In addition, the inverse property name, normally given in parentheses, is also designed to be semantically meaningful and grammatically correct when read from range to domain.

A triple is defined as:

DOMAIN PROPERTY RANGE

The domain is the class (or entity) for which a property is formally defined. Subclasses of the domain class inherit that property. The range is the class that comprises all potential values of a property. Through inheritance, subclasses of the range class can also be values for that property. Again from [CIDb], example 1 illustrates how triples can be extracted from natural language.

(1) Rome identifies the capital of Italy.
    DOMAIN E41   PROPERTY P1   RANGE E1
    E48:Place Name   P1:identifies   E53:Place

   'Rome identifies the capital of Italy.'

The task of the natural language processing tool is to map relevant parts of texts to entities and properties in such a way that triples can be constructed (also known as *Entity and Relationship Extraction*, see [She03]). In a nutshell, the Noun Clauses (NC) *Rome* and *the capital of Italy* are mapped to *Entity 48* and *Entity 53* respectively, themselves subclasses of the domain E41 and range E1 respectively, while the Verb Clause (VC) *identifies* is mapped to *Property P1*. Sections 2 and 3 introduce the CIDOC-CRM standard and the background necessary for processing natural language respectively. Section 4 presents the methodology used to extract triples from texts. The experiments are explained and discussed in section 5 before concluding in section 6.

2. CIDOC-CRM as the documentation standard for Cultural Heritage

CIDOC-CRM, the ISO21127 International Standard under publication as of 06/06/2006, is a Reference Ontology for the Interchange of Cultural Heritage Information. In other words, it serves as a basis for the management of documentation concerning Cultural Heritage, be it a museum collection, an archaeological site or a database of inscriptions. The universality and completeness of this system is increasingly accepted by heritage professionals, who are becoming aware of the existence of such an international and overarching framework. However, the advantages of using a standard are probably still unclear to them, and the burden of managing legacy systems prevents a wide adoption. Furthermore, the compilers of CIDOC-CRM have rightfully chosen a theoretical and supra-institutional perspective, and do not provide application-specific guidance. This does not facilitate the adoption of the system by heritage practitioners. In fact, only a small number of applications may be presently listed [CIDa]. Since the only way to semantic interoperability is the adoption of a common standard for data description, the EPOCH project [EPO] has undertaken the task of creating a tool – named AMA (see [EPO] under RESEARCH and AMA) – to map compatible data structures to CIDOC-CRM. This approach is in our opinion the only feasible one. Firstly, given *n* heritage management systems, it requires the definition of *n* mappings, while a 1-to-1 mapping among them would have required n*(n-1) asymmetric

mappings, and a 1-to-1 mapping to a standard would have required 2n asymmetric mappings. Secondly, since the substance of heritage information is largely the same, the mapping universe AMA will hopefully create will ultimately be a learning system, where new users greatly benefit from the work of previous researchers who already solved most of the problems arising from mapping. It is also possible that in the future, when much information on the mappings will have been acquired, the system may become an intelligent one and suggest solutions basing on the knowledge base accumulated in previous work. Thirdly, this approach solves the problems related to legacy archives, which do not need to be converted to CIDOC-CRM to become interoperable: the data system may remain the same and be used as such for routine work (which in our opinion as yet takes 90% of the time, if not more) and become interoperable via a mapping on-the-fly when this functionality is requested. Still, to achieve full interoperability, there remains the problem of the different language used for data representation: usually data are described and stored in the owner's mother language that creates a barrier to operate with similar databases containing information written in a different language. This obstacle may be circumvented, although not fully eliminated, with the use of multilingual thesauri containing the most significant domain-specific terms. A preliminary expedition in this complex area is being undertaken by EPOCH as well. For the scope of the present paper, the work of AMA is paramount, because it will eventually guarantee the availability of CIDOC-CRM encoded data even when they are stored with a proprietary structure, provided that the task of mapping such a structure to CIDOC-CRM has been accomplished. Conversely, AMA will benefit from the results of the present investigation because it will provide an additional benefit to the list of those deriving from performing the mapping task.

3. Natural Language Processing

Information Extraction (IE) is concerned with the extraction of useful information from text by first using Natural Language Processing (NLP) techniques to get structural information. Figure 1 illustrates the kind of information that can be extracted from example 1. In the remaining of this section we review in turn each element making up the parsed tree of figure 1.

Lemma At the very bottom of the tree we find the lemmas of the words making up example 1, i.e. words which have not been transformed morphologically (e.g. *identify*). Lemmas are more useful for semantic analysis, since they can be looked up directly in a dictionary or thesaurus.

Part-Of-Speech Part-Of-Speeches (POSs) are the grammatical categories of each (inflected) word in a sentence. Some relevant categories for our purpose are IN (preposition or subordinating conjunction), DT (determiner), NN (noun,
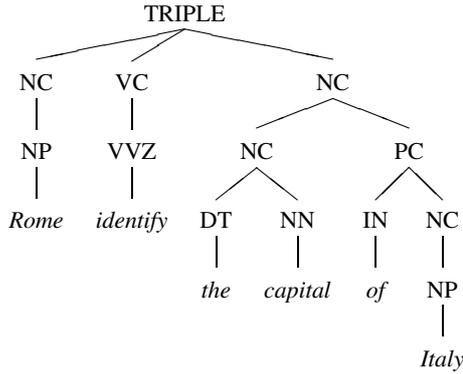
TRIPLE

(Figure: tree diagram)

TRIPLE
- NC → NP → *Rome*
- VC → VVZ → *identify*
- NC:
  - NC:
    - DT → *the*
    - NN → *capital*
  - PC:
    - IN → *of*
    - NC → NP → *Italy*

Figure 1: *Linguistic analysis of example 1.*

singular or mass), NNS (noun, plural), NP (proper noun, singular), NPS (proper noun, plural), V_ (verbs).

Clause A clause is a coherent whole of POSs or other clauses. For our purpose, the relevant clauses are NC (noun clause), VC (verbal clause) and PC (prepositional clause). Clauses are built using phrase structure rules, such as:

$$NC \rightarrow \begin{array}{cc} DT & NN \\ [\text{The}] & [\text{capital}] \end{array}$$

$$VC \rightarrow \begin{array}{cc} VBZ & VVN \\ [\text{is}] & [\text{identified}] \end{array}$$

$$PC \rightarrow \begin{array}{cc} IN & NC \\ [\text{by}] & [\text{Rome}] \end{array}$$

Synonymy and Hypernymy Synonyms are words with similar meanings. A hypernym is a word that is more generic than a given word. Only verbs and nouns can have hypernyms. For example, *entity* is an hypernym of the word *person*. This is similar to the notion of subclasses in CIDOC-CRM. In example 1, E41 and E1 are hypernyms (superclasses) for E48 and E53 respectively. WORDNET [WOR] is the lexical database used for that purpose.

Semantic Association When two words (or group of words, i.e. phrase) tend to co-occur in documents, we can assume that they are semantically related. One way of measuring semantic association is called *Pointwise Mutual Information* (PMI) [CH89]. PMI between two phrases is defined as:

$$\log_2 \frac{prob(ph_1 \text{ is near } ph_2)}{prob(ph_1) * prob(ph_2)} \quad (2)$$

PMI is positive when two phrases tend to co-occur and negative when they tend to be in a complementary distribution. PMI-IR refers to the fact that, as in Information Retrieval (IR), multiple occurrences in the same document count as

just one occurrence: according to [TL03], this seems to yield a better measure of semantic similarity, providing some resistance to noise. Computing probabilities using hit counts from IR, this yields to a value for PMI-IR of:

$$\ln \frac{N * (hits(ph_1 \text{ NEAR } ph_2) + 1/N)}{(hits(ph_1) + 1) * (hits(ph_2) + 1)} \quad (3)$$

where N is the total number of documents in the corpus. Smoothing values ($1/N$ and 1) are chosen so that PMI-IR will be zero for words that are not in the corpus, two phrases are considered *NEAR* if they co-occur within a certain distance of each other, and $\log_2$ has been replaced by ln, since the natural log is more common in the literature for log-odds ratio and this makes no difference for the algorithm.

4. Methodology

Figure 1 suggests that all pairs of NC separated by a VC (and possibly other elements) are potentially valid CIDOC-CRM triples. To validate the triples, we must first make sure that the predicate is relevant by extracting the main verb of the verbal clause (VC) and see if its meaning is similar (synonym) to at least one of the CIDOC-CRM properties. For example, it is possible to use the verb *describe* instead of *identify*. Once a set of possible properties is identified, we must verify if the noun clauses (NC) surrounding the property are related to the DOMAIN and the RANGE of that property. To establish the relation, the first step is to identify the semantics of each NC clause. For English, a good indicator of the NC semantics is the rightmost NN in the clause, excluding any attached PC. The rightmost NN is usually the most significant: for example, in the NC *the museum artifact*, the main focus point is *artifact*, not *museum*. In figure 1 the rightmost NN of *the capital of Italy* is *capital* (excluding the attached PC); this tells us that we are dealing with an object of type *capital*. The second step is to see if the type is a subclass of the DOMAIN or RANGE. Because *entity* (E1) is a hypernym of *capital*, then we conclude that the clause *the capital of Italy* is a subclass of E1:CRM Entity. What if the NC has no NN? This means that the clause is made up of at least one proper noun (*Rome*). To establish the type of a proper noun, we use the Web as corpus and *semantic association* as described previously. We compute how similar the word *Rome* is to each of the CIDOC-CRM classes and choose the most similar as being the type of *Rome* (proper nouns are also looked up in WordNet [WOR]). This gives the following triple:

E41:Appellation		E1:CRM Entity
Rome	P1:identifies	*the capital of Italy*
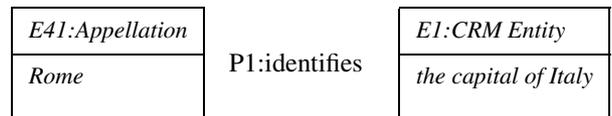
In the remaining of this section we examine the practical details of such a method.

POS tagging and NP chunking POS tagging and NP chunking are combined in one single operation. The method relies on large annotated corpora and statistical machine learning. POS tagging can achieve accuracy as high as 96%. Chunking is the process of grouping POSs in bigger constituents called *clauses*, as previously defined. We have used the freely available and trainable TreeTagger [Sch95], where POS tagging and chunking is available for English and German. We are in the process of creating a tagger and chunker for French.

WordNet: Synonymy and Hypernymy (SH) WordNet is a lexical reference system, developed by the university of Princeton. Its design makes the use of dictionaries more convenient. We have used the Prolog interface. WordNet is based on a concept called *synsets*, also known as synonym sets. A synset is a group of words connected by meaning. Only words of the same part of speech can belong to the same synset. A synset ID is assigned to every word and only words in the same synset have the same synset ID. As one word can have several meanings, it can belong to more than one synset. Then, the word is assigned several entries in the Prolog database, and each entry has a different synset ID assigned. This way we can extract the synonyms of verbs (or properties) and hypernyms of nouns (classes).

PMI: Assigning a class to a proper noun We have used the hit counts provided by the Yahoo [YAH] search engine to compute formula 3, where N is the approximative size of the Yahoo index, hits(ph_1) and hits(ph_2) are simple search while hits(ph_1 NEAR ph_2) is the number of hits returned by Yahoo for a simple conjunctive search ph_1 AND ph_2.

Triple extraction: a walk-through The extraction of CIDOC-CRM triples from text involves mainly the following operations:

1. Text cleaning. The input must be raw text, that is text with no extra tags. Punctuations and special symbols are allowed and, although the system provides some tolerance to grammatical, syntactical and spelling errors, well-formed texts are preferable.
2. Tokenization and POS tagging. Tokenization is the process of splitting the text in individual words or symbols to be POS tagged. POS tagging assigns a POS and a stem (if known) to each token in the form (WORD POS STEM).
3. Clause chunking and pruning. The chunking process assigns clause tags in the form <TAG>...</TAG>, resulting in numerous clauses, which are pruned to the most relevant for our purpose, i.e. NC and PC.
4. NC regrouping. All contiguous NC are regrouped into a single NC ($NC \rightarrow NC+$) and prepositional clauses (PC) following a NC are removed (to get rid of irrelevant subordinate NC clauses).
5. Intermediate triples (IT) creation. Intermediate triples are all *<NC> DOMAIN </NC> PROPERTY <NC> RANGE*

</NC> patterns found in the data. The PROPERTY correspond to the rightmost verb between the domain and the range. They are considered intermediate because they may not correspond to any CIDOC-CRM pattern in the end, given the nature of the verb and NCs. The format of the intermediate triples is (D = DOMAIN, P = PROPERTY, R = RANGE): pred('D_WORD', 'D_STEM':'D_POS', 'P_WORD', 'P_STEM':'P_POS', 'R_WORD', 'R_STEM':'R_POS', '[D] P [R]'). The D_WORD and R_WORD are the rightmost NN, NP or PP (in that order) found in the D and R, respectively. The P_WORD is the rightmost verb in P. The _STEM and _POS are the respective stem and part-of-speech of these words. Finally, in the case the DOMAIN or RANGE is a proper noun (NP), the respective stem is replace by one of the CIDOC-CRM classes according to the PMI measure (or hypernyms found in WordNet).

6. Referent resolution. If a DOMAIN or RANGE is a personal pronoun (i.e. POS = PP), it is replaced by the domain or range of the previous intermediate triple.
7. Final triple (FT) creation. Each intermediate triple is processed to see if they can be matched to a valid CIDOC-CRM triple.

For example 1, this translates as:

1. 'Rome identifies the capital of Italy.'
2. (Rome NP Rome) (identifies VVZ identify) (the DT the) (capital NN capital) (of IN of) (Italy NP Italy) (. SENT .)
3. <NC>Rome Rome NP</NC> identifies identify VVZ <NC>the the DT capital capital NN</NC> <PC>of of IN <NC>Italy Italy NP </NC></PC>. . SENT
4. <NC>Rome Rome NP</NC> identifies identify VVZ <NC>the the DT capital capital NN</NC>. . SENT
5. pred('Rome', 'inscription':'NN', 'identifies', 'identify':'VVZ', 'the capital', 'capital':'NN', '[Rome Rome NP] identifies identify VVZ [the the DT capital capital NN]').
6. No referent resolution
7. IT D:[Rome Rome NP]
   IT P:identifies identify VVZ
   IT R:[the the DT, capital capital NN]
   SH D:[rome, location, group, entity, city, appellation]
   SH P:[identify]
   SH R:[capital, entity, location]
   FT D:[e41:Appellation][Rome]
   FT P:p1:identifies
   FT R:[e1:CRM Entity][the capital]
   FT D:[e41:Appellation][Rome]
   FT P:p1:identifies
   FT R:[e53:Place][the capital]

5. Experiments

We have conducted two experiments. In the first experiment (5.1), we collected all one hundred forty-four examples of triples provided in the CIDOC-CRM documentation. In the

the capital of Italy (E53) is identified by (P1) Rome (E48)
www.cidoc.icom.org (E51) has type (P2) URL (E55)
silver cup 232 (E22) consists of (P45) silver (E57)
chess set 233 (E22) has number of parts (P57) 33 (E60)
height of silver cup 232 (E54) has value (P90) 226 (E60)
height of silver cup 232 (E54) has unit (P91) mm (E58)
Mozart's death (E69) was death of (P100) Mozart (E21)
Late Bronze Age (E4) finishes (P115) Bronze Age (E4)
Early Bronze Age (E4) starts (P116) Bronze Age (E4)
Scotland (E53) borders with (P122) England (E53)

Table 1: *Triples from the CIDOC-CRM documentation [CIDb]*

second experiment (5.2), we have extracted triples from a text describing the medieval city of Wolfenbüttel.

5.1. CIDOC-CRM examples

Table 1 shows a few examples provided in the CIDOC-CRM documentation. In these examples, there were 1965 words and 144 sentences. From this we extracted 149 intermediate triples and 184 final triples. The system has generated at least a final triple for 46 sentences, from which:

- 11 represents a suitable match (DOMAIN PROPERTY RANGE), if we consider the selection of a subclass of DOMAIN or RANGE as acceptable, since the system is being more specific than necessary;
- 29 had the right property, although many mismatches were due to the many similar property sharing the verb 'have' and 'be';
- 21 DOMAINs or RANGEs were being less specific (i.e. a superclass) than the true class;
- 15 DOMAINs or RANGEs were more specific (i.e. a subclass) than the true class.

5.2. Extracting triples from free text

The following experiment shows the result of extracting triples from a textual description of the medieval city of Wolfenbüttel. The document was 3922 words long with 173 sentences. The system extracted 197 intermediate triples and 79 final triples. Table 2 shows a few processing steps for the following fragment of text:

> Lange Herzogstrasse is Wolfenbüttel main shopping area. The street's particular charm lies in its broad-faced half-timbered buildings, historic merchant's houses; their central gables still retain the distinctive hatches through which goods could be hoisted up to the attics for storage.

IT	D1	[Lange Herzogstrasse]
IT	P1	is
IT	R1	[Wolfenbüttel's main shopping area]
SH	D1	[herzogstrasse,attribute]
SH	P1	[be]
SH	R1	[area, entity, location]
IT	D2	[The street's particular charm]
IT	P2	lies in
IT	R2	[its broad-faced half-timbered buildings]
SH	D2	[attribute, charm, entity, language, object]
SH	P2	[consist]
SH	R2	[activity, building, creation, creation, entity, event, object]
FT	D2	[e13:Attribute Assignment]
FT	P2	p9:consists of
FT	R2	[e7:Activity]
FT	D2	[e13:Attribute Assignment]
FT	P2	p9:consists of
FT	R2	[e65:Creation Event]
FT	D2	[e13:Attribute Assignment]
FT	P2	p9:consists of
FT	R2	[e5:Event]
IT	D3	[their central gables]
IT	P3	still retain
IT	R3	[the distinctive hatches goods]
SH	D3	gable
SH	P3	retain
SH	R3	good
IT	D4	[the distinctive hatches goods]
IT	P4	could be hoisted up to
IT	R4	[the attics]
SH	D4	good
SH	P4	hoist
SH	R4	attic

Table 2: *A few triples extracted from free text.*

Synonyms and hypernyms are also shown for domains (D), properties (P) and ranges (R). For example, *attribute* is the result of looking for the highest PMI-IR value for the proper noun *Herzogstrasse*, *consist* is a synonym for *lie*, and *entity, location* are hypernyms of *area*. In each case, we extracted from WordNet the synonyms and hypernyms of the three most common uses for each word (verb, noun). In terms of processing speed, steps 1 to 5 (in Perl) take no more than a few seconds, unless we must look proper noun on the Web; using the Yahoo API interface, each PMI-IR computation takes approximately one and a half minute. For steps 6 and 7 (in Prolog), the treatment of intermediate triples, we must allow almost 2 minutes for each intermediate triple to be fully processed.

5.3. Discussion

It is difficult to have a comprehensive evaluation of the system through standard metrics (precision, recall), since there is no benchmark for this type of analysis. A good benchmark would be a CIDOC-CRM human-annotated text. Yet we can give some evidence of the performance of the system. In the first experiment, although there were only 11 perfect matches, many more had at least a suitable property, and a few of these had either a domain or a range which was appropriate. An important cause of mismatch is that many properties are expressed through the verbs *be* or *have*, for which the system cannot make a distinction; extracting more information adjacent to the verbal clause should improve the accuracy of the system. Last but not least, the 149 intermediate triples offer a good fall-back in case the recall of final triples is too low. In the second experiment, we have collected 79 final triples from a 173 sentences long document describing buildings and places of interest in a medieval city. The data was relatively clean, although punctuation was heavily used throughout the document, confusing the chunker. Despite the fact that recall and accuracy appear to be low, there is no doubt that a system like this gives a head start to anyone wishing to build a collection using the CIDOC-CRM ontology. A first pass in the documentation gives a good idea of what the textual documentation is about. However, a fuller interpretation will often involve combining many triples together to form paths. Because of time restriction, we have elected to process the three most common meanings of each word that we looked up in WordNet (avoiding the need to manually pick the right meaning among many); this may have the side effect of lowering accuracy. Speed was not an issue without access to the Web, not an absolute necessity if we have a good thesaurus for proper nouns. Finally, we have tuned the CRM to analyse impressions of a city, which is not a domain for which the CRM is optimally intended. We conjecture that texts about museum catalogues would have yielded better results.

6. Conclusion and Further Work

We have presented a method for extracting CIDOC-CRM triples using language technology. The tool presented exploit the propositional nature of CIDOC-CRM triples and use pattern matching approach based on the output of a phrasal chunker for noun and verbal phrases. The result is a flexible tool that gives a good approximation of the semantic nature of text, from a CIDOC-CRM angle. It can be readily adapted to other languages. The results of the experiments are modest but worth further investigating. The most pressing areas of research includes domain specific thesauri such as [STA, MDA] and discriminatory methods for properties and entities, including common linguistic constructions that do not match the expected [entity, property, entity] pattern. The system can be paired with a more formal mapping method to form a robust translator for diverse digital formats

into CIDOC-CRM triples. Finally, let's not underestimate the positive impact of cleaner textual data on the accuracy of our retrieval system.

References

[CH89]   CHURCH K. W., HANKS P.: Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics* (Vancouver, B.C., 1989), Association for Computational Linguistics, pp. 76–83. 3

[CIDa]   Applications and uses for CIDOC CRM. http://cidoc.ics.forth.gr/uses_applications.html. 2

[CIDb]   CIDOC-CRM. http://cidoc.ics.forth.gr/. 1, 2, 5

[EPO]   EPOCH. http://www.epoch-net.org/. 2

[MDA]   MDA Archaeological Objects Thesaurus. http://ww.mda.org.uk/archobj/archcon.htm. 6

[Sch95]   SCHMID H.: Improvements In Part-of-Speech Tagging With an Application To German. *EACL SIGDAT workshop* (1995). 4

[She03]   SHETH A.: Capturing and applying existing knowledge to semantic applications. Invited Talk "Sharing the Knowledge" - International CIDOC CRM Symposium, March 2003. Washington DC. 2

[STA]   The Stanford WordNet Project. http://ai.stanford.edu/ rion/swn. 6

[TL03]   TURNEY P., LITTMAN M.: Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems 21*, 4 (2003), 315–346. 3

[WOR]   WORDNET. http://wordnet.princeton.edu/. 3

[YAH]   YAHOO. http://www.yahoo.com/. 4