Exploiting glossaries for automatic terminology processing
Le An Ha
Research Group in Computational Linguistics,
Research Institute in Information and Language Processing
University of Wolverhampton, Stafford Street, Wolverhampton
WV1 1SB, UK
L.A.Ha@wlv.ac.uk

Abstract
This paper analyses a valuable but forgotten resource in automatic terminology processing (ATP): glossaries. It argues that glossaries are widely available, especially on the Internet, and that they contain valuable terminological knowledge which can be exploited by automatic procedures. The empirical analysis of a set of glossaries collected from the Internet substantiates these arguments. The paper also presents a method to extract knowledge patterns from glossaries. An evaluation is then performed showing the usefulness of the extracted patterns in ATP. In two experimented domains, the improvements are 5% and 16% over f-measures respectively. The paper concludes that glossaries should be further studied and exploited.

Keywords
Automatic terminology processing, automatic term extraction, pattern extraction, pattern heuristics.

1. Introduction
Automatic terminology processing (ATP) concerns the deployment of automatic or semi-automatic procedures to build, maintain, and exploit terminologies. Although generic ATP methods have been proposed ([10]), adapting them to different domains remains a challenge. To address this problem, it has been argued that glossaries (lists of the most important terms relating to a specific domain, together with their brief definitions or explanations) can be used. This is because glossaries are widely available and contain domain-specific terminological knowledge that, if extractable, can be used by ATP engines.

This paper discusses which terminological knowledge can be extracted from a glossary, and how. We first discuss what glossaries are, their features, and a sample set of 7 glossaries collected from the Internet (Section 2). The types of terminological knowledge which glossaries contain and how they can be retrieved (Section 3) are then discussed. An evaluation will show the extent to which the extracted knowledge is useful for ATP (Section 4). Conclusion will be found in Section 5.

2. Glossaries and their features
2.1 Definitions and usages
In contrast to other terminological resources, there is a lack of studies on glossaries in terminology processing literature. More often, authors discuss dictionaries or encyclopaediae ([15], [16], and [19]), possibly because they consider glossaries to be similar to them.

Definitions of a glossary generally agree that it is a list of technical terms along with their brief explanations, and that glossaries can be used for alternative purposes such as a reference point of a book, a common terminology for internal communication of a company, or a place where explanations of jargon used on a website can be found.

2.2 Glossaries in the information era
In this information and knowledge era, the general public constantly exploits increasingly available resources for their own needs using the Internet and the World Wide Web (WWW). Just as with books, there may be several terms on a website with which some readers are not familiar, and a short explanation is needed. Recognising this need, website authors put glossaries onto their websites and enhance them with features provided by the Internet and WWW such as hyperlinks, multimedia presentations, and search facilities. Search engines (such as Google) have developed search features which exploit available glossaries to allow users to find definitions of words and phrases; this confirms that there are both supply of and demand for glossaries.

2.3 Collecting glossaries from the Internet
To confirm the hypothesis that glossaries contain valuable terminological knowledge which can be extracted and used in ATP, we collected a set of glossaries from various domains from the Internet to be used for empirical analysis and to design algorithms to extract terminological knowledge. To do this, firstly we searched Google for the keyword “glossary”, and obtained the glossaries from the first 100 results. In this study, we discuss a set of 7 glossaries covering 7 different domains. Domains and descriptions of the selected ones are shown in Table 1.

Glossaries on the WWW are presented in different ways and formats, varying from plain text only to those that use every available hypertext features. The hypertext markup language ([21]) does provide a set of html tags to be used to mark terms and definitions; they are <DL>, <DT>, and <DD>. When this tagset is used, terms are usually highlighted in bold and their definitions indented. However, not all compilers of Internet glossaries are aware of, or want to use, this tagset. In our sample set, three glossaries use other html tags to identify terms and explanations.
### 2.4 Features of glossaries

Empirical analysis shows that glossaries’ features can be divided into two categories: i) essential features: required in order for something to be considered a glossary; ii) supplementary features: used to enrich glossaries. Essential features of a glossary include: i) a list of terms; ii) a short description attached to each term (which will be referred to as a gloss); iii) a method to quickly search for entries. The list of terms contains entries which the authors think is important, or worthy of inclusion; the numbers of entries in our set vary from 121 (CITIZEN) to 2641 (WEATHER). Each entry in a glossary is followed by a gloss (short description). Entries in glossary are sorted alphabetically help searching. Supplementary features of glossaries include: i) cross references: provide reader with references to other relevant terms, and ii) multimedia presentation: used to present information using audio and animated visual effects. In our sample set of glossaries, all but WEATHER provide hyperlinked cross references to other relevant terms. Only CHEMISTRY has multimedia presentation.

The glosses in a glossary constitute its most important parts. A gloss (of an entry) provides description, explanation, or any information the author thinks may help readers to understand the entry quickly. The following extract is an example of glosses.

**absolute temperature.** Temperature measured on a scale that sets absolute zero as zero. In the SI system, the Kelvin scale is used to measure absolute temperature. (CHEMISTRY)

Generally speaking, a gloss is different from a definition found in a dictionary or encyclopaedia. Written by domain experts rather than lexicographers, they tend to be more informal. A gloss can also be considered a summary of information that would provide readers with concise knowledge of the term ([8]). A summary of various word statistics on glosses can be found in Table 2. Average numbers of sentences per gloss vary from 2.4 to 3.3; average numbers of words per gloss: 26.8 - 66.6.

### Table 1: Domains and descriptions of the collected glossaries

<table>
<thead>
<tr>
<th>Glossary</th>
<th>Description</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAVA</td>
<td>Java Reference</td>
<td>Java</td>
</tr>
<tr>
<td>WEATHER</td>
<td>National Weather Service Glossary</td>
<td>Weather</td>
</tr>
<tr>
<td>CANCER</td>
<td>CancerhelpWebsite glossary</td>
<td>Cancer</td>
</tr>
<tr>
<td>UNICODE</td>
<td>The Unicode Standard, Version 4.0</td>
<td>Unicode</td>
</tr>
<tr>
<td>CITIZEN</td>
<td>Glossary and Acronym of the U.S. US Citizen and Immigration services Citizenship</td>
<td>CHEMISTRY General Chemistry Glossary Chemistry</td>
</tr>
<tr>
<td>WATER</td>
<td>Water Science Glossary of Terms</td>
<td>Water science</td>
</tr>
</tbody>
</table>

### Table 2: Number (#) of sentences and words in glosses

<table>
<thead>
<tr>
<th>Glossary</th>
<th>Total # of sentences</th>
<th>Aver # sentences/gloss</th>
<th>Number of words</th>
<th>Aver # words/gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAVA</td>
<td>684</td>
<td>2.67</td>
<td>6860</td>
<td>26.80</td>
</tr>
<tr>
<td>WEATHER</td>
<td>6353</td>
<td>2.40</td>
<td>78827</td>
<td>29.85</td>
</tr>
<tr>
<td>CANCER</td>
<td>3515</td>
<td>2.70</td>
<td>34229</td>
<td>54.19</td>
</tr>
<tr>
<td>UNICODE</td>
<td>858</td>
<td>2.47</td>
<td>10095</td>
<td>29.09</td>
</tr>
<tr>
<td>CITIZEN</td>
<td>393</td>
<td>3.25</td>
<td>8053</td>
<td>66.55</td>
</tr>
<tr>
<td>CHEMISTRY</td>
<td>3010</td>
<td>2.88</td>
<td>35159</td>
<td>33.68</td>
</tr>
<tr>
<td>WATER</td>
<td>458</td>
<td>2.99</td>
<td>61960</td>
<td>40.46</td>
</tr>
</tbody>
</table>

3. Exploiting glosses

Given that glosses are used to explain the meaning of their entries and to provide important information about them, we first discuss several studies of definitions in the field of terminology processing. We then discuss methods to extract useful terminological knowledge from glosses.

#### 3.1 The study of definitions in terminology processing literature

According to [2], [9], [15], and [20], the classic formula for a definition is $X = Y + \text{distinguishing characteristics (differentia)}$, in which $X$ is the entry, and $Y$ is a genus$^2$ term superordinating $X$. The differentia differentiates $X$ from other concepts in the domain. Swales ([20]) has argued that the definition formula is often realised using a set of linguistic patterns; most of these patterns occur in our selected glossaries.

#### 3.2 Pre-processing glosses

Parsing technologies allow us to analyse glosses quickly without a great deal of errors. Parsers such as that of [3] provide a reasonably accurate shallow syntactical analysis of a sentence. Using the output of [3], we can analyse the collected glosses in terms of sentence structure as well as the head words of these structures.

We use the parser to process the selected glossaries. The parser’s outputs have proved to be sufficient, apart from some consistent errors fixed by post-processing rules. Analysing glosses using parser’s outputs also provides an indication of parser’s performance in ATP tasks.

Using the parser output, the genus terms (in the definition formula) can be retrieved. The genus terms are located in the ‘first sentence’ of a gloss. If the ‘first sentence’ of a gloss is an NP, the genus is its head (e.g. *Temperature measured on a scale that sets absolute zero as zero: temperature*). If the ‘first sentence’ is a complete sentence, the genus is often the head of the argument of the copular verb (*Visible light is electromagnetic radiation with a wavelength between 400 and 750 nm: radiation*). Table 3 presents the ten most used genus terms extracted from selected glossaries.

---

1 We borrow the term “gloss” from WordNet ([6]) to describe the information attached to a term in a glossary. Originally, glossary meant a collection of glosses, and glosses were notes made in the margins or between the lines of a book ([14]).

2 The terms “genus” and “differentia” are borrowed from [2].
Table 3: Ten most popular genus terms extracted from glossaries

<table>
<thead>
<tr>
<th>Glossary</th>
<th>Genus terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAVA</td>
<td>keyword, protocol, method, item, system, class, language, unit, type, definition</td>
</tr>
<tr>
<td>WEATHER</td>
<td>system, model, time, area, wind, term, cloud, center, product, instrument</td>
</tr>
<tr>
<td>CANCER</td>
<td>cancer, operation, treatment, cell, lymphoma, drug, doctor, tube, substance, disease</td>
</tr>
<tr>
<td>UNICODE</td>
<td>character, acronym, synonym, standard, sequence, name, system, script, set, collection</td>
</tr>
<tr>
<td>CITIZEN</td>
<td>alien, category, limit, child, provision, number, immigrant, person, public, law</td>
</tr>
<tr>
<td>CHEMISTRY</td>
<td>substance, compound, reaction, example, unit, element, change, prefix, acid, process</td>
</tr>
<tr>
<td>WATER</td>
<td>water, process, substance, rock, term, unit, measure, amount, feature, system</td>
</tr>
</tbody>
</table>

3.3 Differentiae, use of verbs, and knowledge patterns in glosses

In glosses, often there is no differentiae element (as in the classic formula of definition). Rather, there is an explanation why the term is important. The following example illustrates this observation.

**Alpha-Fetoprotein (AFP)**. Substance found in the bloodstream of some men with testicular cancer. The level rises when the cancer is growing and falls when the cancer is shrinking. … (CANCER)

In this example, it can be argued that there are many substances which can be found in the bloodstream of men with testicular cancer. Thus, the first ‘sentence’ explains why AFP is important in Cancer rather than trying to distinguish it from other substances. It provides also a connection from AFP to other important concepts in the domain (i.e. testicular cancer). Consider another example:

**Xenylamine**. Chemical which has been found to cause bladder cancer. (CANCER)

In this example, the most important fact about Xenylamine is that it is found to cause bladder cancer; its connection with the domain. Following examples reinforce the view that glosses of an entry justify its inclusion, often by establishing the entry’s connection with the domain.

**Case**. A Java keyword that defines a group of statements to begin executing if a value specified matches the value defined by a preceding switch keyword. (JAVA)

Here, the gloss is also a true definition stating the differences between case and other Java keywords.

**Aqueduct**. a pipe, conduit, or channel designed to transport water from a remote source, usually by gravity. (WATER)

The connection between Aqueduct and the WATER domain is that an aqueduct transports water.

**Certificate of Citizenship**. Identity document proving U.S. citizenship. Certificates of citizenship are issued to derivative citizens and to persons who acquired U.S. citizenship (see definitions for Acquired and Derivative Citizenship). (CITIZENSHIP)

The connection between Certificate of Citizenship and the US Citizenship domain is that Certificate of Citizenship is an identity document, Certificate of Citizenship proves U.S. citizenship, and Certificate of Citizenship is issued to derivative and acquired citizens.

In the majority of cases, relations between the entry and the domain are explicitly stated using verbs such as “contain” (CHEMISTRY), “cause” (CANCER), and “define” (JAVA). Empirical observation suggests that such verbs, whilst varying across different domains, are used repeatedly within a glossary, and thus retrievable [6].

It can be argued that these significant verbs are, in fact, the central parts of the knowledge patterns which signal the important knowledge in a field, such as “A CONTAIN B” in the domain of Chemistry:

- **acid**: a compound containing detachable hydrogen ions;
- **alloy**: a mixture containing mostly metals;
- or “A STOP B” in the domain of Cancer:
  - **Anaesthetic**: Drug which stops feeling, especially pain;
  - **Aminoglutethamide**: Drug used to treat breast cancer which stops the Adrenal Gland from making sex hormones. …

The notion of knowledge patterns in ATP has already been discussed in various studies ([1], [5], [11], [15]), although different terms may be used instead of knowledge patterns. The study of knowledge patterns assumes a general meaning: a linguistic pattern which expresses important knowledge in the domain. We shall focus on patterns whose anchors are verbs, for example “X IS_A compound”, “X CONTAIN ring”, “X CONTAIN Y”, and “X TREAT Y”.

3.4 Extracting and scoring knowledge patterns from glossaries

3.4.1 Pattern extraction in NLP

Pattern extraction is an interesting topic in NLP, as patterns are a means to extract further information. A pattern extraction method often has two components: a pattern heuristic and a pattern scoring method. A pattern heuristic is needed in order to identify pattern candidates. Once identified, pattern candidates are assigned scores so that significant patterns have a greater effect on the intended tasks. Relevant works that propose pattern heuristics and scoring methods include [12], [13], [17], and [18].

3.4.2 The proposed pattern heuristic

Similar to [17] and [18], we concentrate on subject–verb–argument patterns which often express important relations. As subjects (the entries being described) in sentences in glossaries are often omitted, it is safe to concentrate on the verb–argument parts of knowledge patterns. We propose a pattern heuristic that will capture patterns from glossaries at three levels of detail as follows:

**VERB + NP** (the verb is followed by an NP)
VERB + TERM (the verb is followed by a TERM found in the glossary)
VERB + head (the verb is followed by a specific head word of an NP)

The three levels are intended to capture patterns at different levels of detail, leaving the pattern scoring method to assess their significance. The first two pattern heuristics are similar to those proposed in the literature. The additional pattern heuristic (VERB + NP’s head) is intended to capture patterns of general verbs (be, have, etc.) which may otherwise be overlooked by other heuristics. To illustrate this ability, consider the following context: “is a compound”, from which other pattern heuristics may suggest only two pattern candidates: “be NP” and “be <TERM>”, both considered too general for the domain of Chemistry. In this case, the third level is used to suggest the pattern candidate: “be compound”. The output of the parser is to identify VPs and the verb’s arguments. Following is an example of how the pattern heuristic works:

**Burkitt’s Lymphoma.** Burkitt’s Lymphoma is a rare and special type of lymphoma that is usually treated with combination chemotherapy.

From this sentence, the parser returns:

(51 (S (NP (NP (NN Burkitt) (POS ’s)) (NNP Lymphoma))

(S (VP (AUX is)

(NP (NP (DT a) (ADJP (JJ rare) (CC and) (JJ special)) (NN type))

(P (IN of) (NP (NN lymphoma))))

(SBAR (WHNP (WDT that))

(S (VP (AUX is) (ADVP (RB usually)))

(VP (VBN treated) (PP (IN with)

(NP (NN combination) (NN chemotherapy))))))))

which contains three VPs. For the first VP (“is a rare type of … treated with combination chemotherapy”), the algorithm suggests two patterns: “BE lymphoma” and “BE NP”. This is because lymphoma is identified as the head of the NP “a rare and special type of lymphoma that is usually …”. The second VP (“is usually treated with combination chemotherapy”) does not produce any pattern candidate. For the third VP (“treated with combination chemotherapy”), the algorithm discovers three more patterns: “TREATED WITH chemotherapy”, “TREATED WITH <TERM>”, and “TREATED WITH NP”.

We call patterns which have <TERM> as their arguments (such as CONTAIN <TERM>) “binary patterns”. Patterns which have specific words as their arguments (such as “BE lymphoma”) are called “unary patterns”. It can be said that two types of patterns reflect two types of relations: relations between two individual terms and relations between individual terms and the whole terminology (the domain).

### 3.4.3 Assigning scores to pattern candidates

In ATP, it is important to assign scores to pattern candidates so that significant patterns have higher scores and, as a result, stronger influence on ATP than insignificant patterns. Several scoring methods have been experimented with and among them, the following formula has proved to be the best way to score patterns:

$$S_{Ra}(p_i) = \frac{F(p_i)}{Fr(p_i)} \cdot \log(F(p_i))$$

In this formula, $F(p_i)$ denotes the frequency of the pattern $p_i$ in the glossary and $Fr(p_i)$: the frequency of $p_i$ in the reference corpus. This scoring method rewards both patterns which occur frequently in the glossary, and patterns which occur frequently in the glossary in comparison to in a reference corpus. This scoring method can be considered similar to those of [13] and [18]. Examples of high-scoring patterns include: “FIGHT <TERM>”, “INCREASE risk” (from the glossary CANCER); “CONTAIN ring”, “DISSOLVED IN <TERM>” (CHEMISTRY).

## 4. Knowledge patterns and ATP

### 4.1 Incorporating knowledge patterns

The extracted knowledge patterns can be considered as semantic information, which has already been used in ATP. A generalised way to incorporate semantic information into the termhood function is to add semantic information scores to it: $FK(t) = F(t) + \alpha_1K_1(t) + \alpha_2K_2(t)$

In this formula, $F(t)$ is the original termhood function (such as frequency and C-value), $K_1(t)$: the score of the semantic information contexts of the term candidate $t$ independent of other term candidates, $K_2(t)$: the score of the semantic information contexts which also involve other term candidates, and $\alpha_1, \alpha_2$: the weights of these scores. In our case, $K_1(t)$ is calculated as:

$$K_1(t) = \sum_{p \in C(t)} S(p)$$

Here, $C(t)$: the set of all instances where a unary knowledge pattern $p$ suggests a relation between the term candidate $t$ and the pattern’s right argument; $S(p)$: the score of the pattern $p$ of the instance (the calculation of this score has been discussed in the previous section); and $K_2(t)$:

$$K_2(t) = \sum_{p \in C_{3}(t)} S(p)F(t)$$

Here, $C_3(t)$ is the set of all instances where a relation between the term candidate $t$ and another term candidate $t'$ suggested by a binary pattern $p$, and $F(t)$ is the termhood score of the term $t'$, which is the right argument of the pattern $p$ in the instance. $K_2(t)$ can also be calculated recursively. $\alpha_1$ and $\alpha_2$ values are assigned by experiments.

3 A termhood score is a score which indicates how likely a term candidate is a term. A termhood function assigns termhood score to a term candidate.
4.2 Evaluation

We choose to evaluate our proposed methodology over two domains: Cancer and Chemistry, whose knowledge patterns were extracted from their glossaries (see Sections 2 and 3). Texts for the two chosen domains were collected from the Internet. The Cancer domain corpus (CanCor) contains 1248 documents (750,000 words); the Chemistry corpus (ChemCor) contains 300 documents (380,000 words).

4.2.1 Gold standard and evaluation metrics

To evaluate the quality of the extracted terms, we compare the outputs (the term candidate lists) provided by different termhood functions. For each domain, we combine three different glossaries to form a final list of terms. The number of glossaries in which a term appears is used as its weight. String matching is used to estimate the total number of terms, their total weights, and the average weight of a term which can be extracted from these two corpora.

For a list of top N term candidates proposed by a termhood function, precision is calculated as the total weight of correct terms (weighted hits) divided by the average weight of N terms in the corpus, recall is the weighted hits divided by the total weight of terms identified using string matching, and F-measure is calculated as usual.

4.2.2 Results

The results show that the average improvements over the baseline termhood function (frequency) of F-measures over six values of N from Cancor and ChemCor are 5% and 16% (statistically significant at p=0.05) respectively. It is observed that knowledge patterns have a greater effect on shorter lists of term candidates than longer ones.

5. Conclusions

This paper discusses a forgotten resource in ATP: glossaries. We have argued that glossaries are increasingly available and used; and that they contain valuable terminological knowledge. A method to extract one type of terminological knowledge presented in a glossary - knowledge patterns – is discussed. The extracted knowledge patterns are then shown to be useful in ATP: it helps increase the performance of automatic term extraction (in term of weighted f-measures) by 5% and 16% respectively over two corpora. These extracted knowledge patterns can also be used in other ATP tasks ([8], [11]).

References