Building Knowledge Base for Reading from Encyclopedia*

Ka Kan Lo and **Wai Lam**
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong,
Hong Kong
{kklo,wlam}@se.cuhk.edu.hk

Abstract

This paper presents a framework using Encyclopedia texts to induce and generalize entities and relations existing in the corpus so that the extracted elements can be used for processing more general text. We begin with the general description of the sense model of the corpus. We then formalize the approach of automatically extracting and structuralizing entities and relations from the corpus.

Introduction

Reading is the universal method of working with texts and a way of acquiring knowledge. In many subfields of AI such as automated reasoning, knowledge representation and reasoning, "knowledge" are the major ingredient in building a machine with plausible reasoning mechanism. By building a coherent set of beliefs, reasoning can then follow. Existing approaches, in which the knowledge experts codify the knowledge and manually or semi-automatically enter into the system, seem to be the most intuitive way to build up a knowledge system from scratch. However, inconsistency in the knowledge from the experts, inconsistency of the background theory and the semantics of the formal logic framework and the worst, the tremendous amount of labor hour in coding knowledge seems to be the bottleneck of making the general knowledge system become a reality. Of course, the most obvious solution is to let the machine "read" by itself, extract the knowledge codified in the texts and gather a coherent set of entities and relations from the texts automatically.

When the machine reads the text, it has to form hypotheses about the relationships of different entities and relations from the texts. Technically, it can be thought of a process modeling a number of relations so that the concept and possible world coded by the text can be represented in a structural way. It is this structural representation, rather than the apparently random tokens in the text, to make the formation of belief, propositions and reasoning possible. Unsupervised learning of the relations from text seems to be a good starting point of the previously stated difficulties in building comprehensive knowledge systems. Current state-of-the-art unsupervised learning methodology has made some progress in this direction (Smith & Eisner 2004; Widdows 2003), but the major problem of overfitting and unbounded error hinders the direct application of this learning approach.

On the other hand, supervised and semi-supervised learning have demonstrated the potential of exploiting the corpus with pre-tagged relations to enable a machine to discover a large set of relations. (Zhao & Grishman 2005; Chen *et al.* 2006) Existing works have shown that using a large data set such as the web corpus is able to obtain a huge set of relations for applications such as web search. However, the relations extracted are not structured and lack the inter-relationship. Also, it is still quite a challenge to extract relations from text due to the inherent complexity of the natural language. These complexities, such as anaphora, quantification, unbounded dependencies and in addition, noise in the language data, make the process of building relations from the text more problematic.

Current NLP research and computational linguistic research provide an insight and formal model in explaining various types of linguistic phenomena. (Bresnan 2001; Pollard & Sag 1994) This explanatory power can provide part of the solution in which the text, in syntactic form, is translated into semantics. In this paper, we demonstrate the feasibility of using the linguistically sophisticated framework, with appropriate training data using an emerging corpus, to build up a set of relations that in a sense, to structuralize the way the machine may perceive the possible world.

This paper is organized as follows: The next section contains the background and related works. It is followed by three sections on the model proposed for building up entities and relations. The final section gives the conclusions.

Background

In this section, we give an overview of the work in acquiring the knowledge for reasoning from text-like resources and state some of the inherent difficulties in approaching the goal

*The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Nos: CUHK 4179/03E and CUHK4193/04E), the Direct Grant of the Faculty of Engineering, CUHK (Project Code: 2050363). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

of machine reading. One approach is to extract general relations and knowledge from Treebank corpus (Schubert 2002) using pattern matching. The corpus is first parsed using a statistical parser trained by the corpus. The resulting parses are matched against some hand-built patterns to extract the predicate-argument relations of the sentences. However, using simple parse trees ignores the inherent complexities of natural languages. Another method is to use psycholinguistic motivated resources such as WordNet to extract and expand the set of relations (Clark *et al.* 2006) to facilitate other inference tasks. Though it also gathers relations and tries to form a set of proposition for reasoning, the target corpus is not the general natural language text. The goal of this work deviates from the goal of machine reading. Also, it is doubtful whether the simple pattern matching in extracting and expanding relations in WordNet can be extended to more general texts in which the genres and writing style vary greatly.

The general goal of complete machine understanding and reading is still far beyond the current theoretical analysis and computational capabilities. Recent development in computational linguistic, availabilities of new types of corpus, linguistic and knowledge resources and advanced in machine learning, however, can have the potential to stretch the envelop of the state-of-the-art development of machine-related language tasks.

In this paper, we advocate that a plausible approach in enabling machines to process texts generally is to "guide them to read". This means that we are neither giving any non-arbitrary reading items to the machine to train them to automatically extract the relations from texts, nor providing a tagged corpus with pre-defined tags for the machine to learn and extract relations. Instead, we are relying on some new type of corpus such as encyclopedia in which the texts are tagged based on the sense relation. By exploiting the linkages between different concepts in the encyclopedia, the machine gradually gathers a set of linkages relating different individuals and properties in which the machine perceives in the possible world, mirrored by the encyclopedia. To enable the machine to "read", rather than extracting relations based on the surface word orders running in natural language text, we have to rely on some sophisticated computational linguistic approach; HPSG (Pollard & Sag 1994) framework is chosen as the major vehicle in converting the sentences into semantic representation due to its wide empirical coverage. Though the resulting group of relations is based on some predefined corpus of encyclopedia, we believe that the relations learnt can be easily applied and extended to the other texts.

Characteristics of Text Corpus

In this section, we explain the mechanism in which the relationships between entities, relations existing in the possible world can be extracted, induced, and generalized from Encyclopedia texts. The generalized entities and relations can then be applied to other more general texts so that the whole set of entities and relations within the possible world as perceived by the reading algorithm can be expanded when more text is being read.

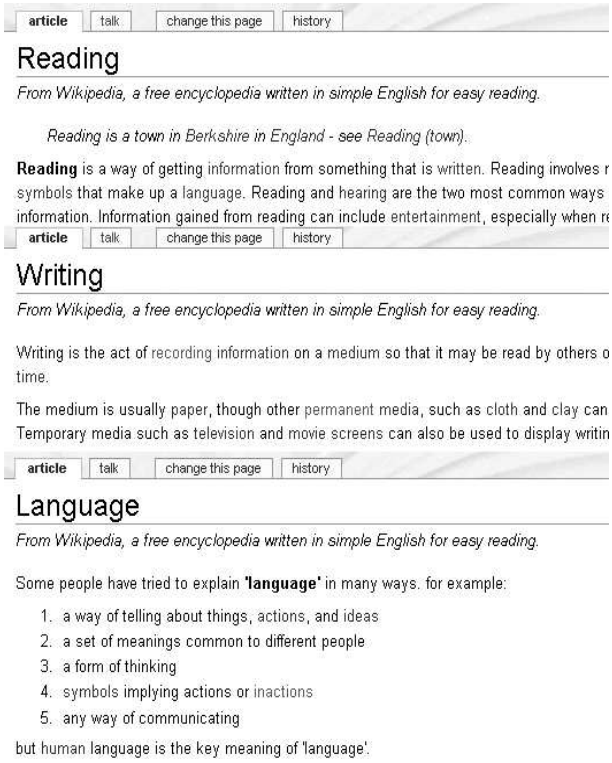

Figure 1: Sample entries extracted from Wikipedia, showing some of the link structures between "reading", "writing" and "language".

Encyclopedia Text

Encyclopedia texts provide hints and suggestions to the relationships between entities and relations to the world. They act as a foundation in which further information between entities and relations can be generalized. This type of texts has the following properties:

1. Partial Sense Linkage: The texts are partially tagged with relationships. Given a sentence within the text corpus, some words within the sentences are linked to the senses in which the words represent. The linked sense may be other essays or paragraphs containing more detailed semantic description of the given word.

2. Linkage Granularity: The links pointing to the targeted senses from a word should contain only that information about the senses, irrelevant information should be avoided.

One recently available online encyclopedia is Wikipedia. The senses are tagged and the link contains only the information about the sense as shown in Figure 1. In this paper, we will focus on the sense linkage inherited from the Wikipedia to induce and generalize entities and relations.

Sense Model

The sense model describes the macroscopic relations between different senses, pages and paragraphs within the

corpus. It also provides the mathematical foundation in describing the mechanism of entities and relations induction.
In the encyclopedia, each page describes a particular concept in which the words represent. Each particular page can be thought of representing the sense of the word. In any pages, there are a number of paragraphs, describing a particular aspect of the senses. In each paragraphs, there are a number of sentences where in some of these sentences, the words within these sentences are tagged with relationships to the other encyclopedia page. These relations can be summarized as follows:

Definition 1: (Sense Model)

Given a sense p_i:

$$\begin{aligned} p_i &= <\{paragraph_{ij}\}, \{insense_n\}> \\ paragraph_{ij} &= \{sentence_{ijk}\} \\ sentence_{ijk} &= <\{word_{ijkm}\}, \{relation_p\}> \end{aligned}$$

For a word $word_{ijkm}$:

$word_{ijkm} = p_m$, if a link exists, pointing from the current sense p_i to the targeted sense p_m. And,

$word_{ijkm} = 0$, otherwise.

$paragraph_{ij}$:The paragraphs within the sense p_i
$sentence_{ijk}$: The sentences within the $paragraph_{ij}$
$word_{ijkm}$: Words within the sentences.
$relation_p$: Relations extracted from the sentence $sentence_{ijk}$. Detailed description on this attribute will be given in next section.
$insense_n$: The set of senses pointing from other senses to this sense.

Example Consider the example in Figure 2, there are three paragraphs within the sense "reading" and ten sense links within these paragraphs.

Let p_1 be the sense of the page "reading".
$sentence_{111} =< \{p_{information}, p_{written}\}, \{relation_1\} >$
0, if otherwise
$sentence_{112} =< \{p_{symbols}, p_{language}\}, \{relation_2\} >$
0, if otherwise
$sentence_{113} =< \{p_{hearing}\}, \{relation_3\} >$
0, if otherwise
$sentence_{114} =< \{p_{entertainment}, p_{fiction}\}, \{relation_4\} >$
0, ifotherwise
$sentence_{121} =< \{p_{paper}, p_{blackboard}\}, \{relation_5\} >$
0, if otherwise
$sentence_{122} =< \{p_{computer}\}, \{relation_6\} >$
0, if otherwise

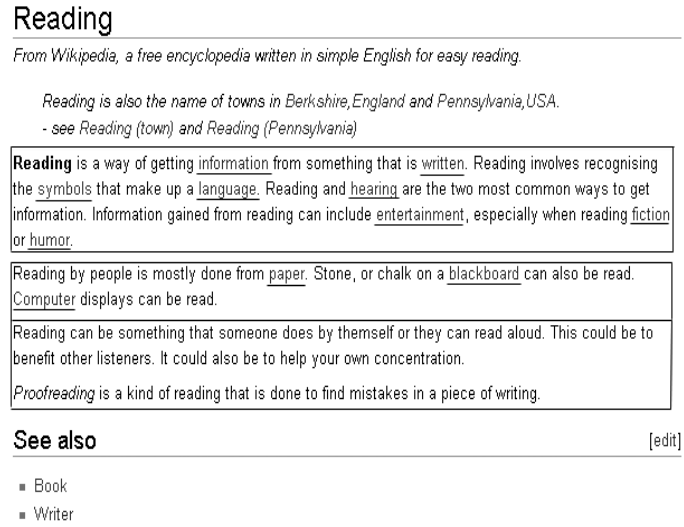

Figure 2: Sense model of the entries of Encyclopedia, showing the relationships between page, paragraphs, sentences and words

For the paragraph:
$paragraph_{11} =$
$< \{sentense_{111}, ..., sentence_{114}\}, \{relation_p\} >$
$paragraph_{12} =$
$< \{sentence_{121}, sentence_{122}\}, \{relation_p\} >$

For the page:
$p_1 =$
$< \{paragraph_{11}, paragraph_{12}, paragraph_{13}, \{insense_n\} >$

Our Proposed Framework

Our proposed framework consist of three main phases: extraction, induction, and generalization.

Extraction

The first stage involves extracting the static relations from the corpus. The relation is static as it only contains the local information of the relation, without linking to the other entities and relations within the corpus.
We adopt the feature structure notation to describe the triple due to the underlying extracting mechanism from linguistic framework. However, any notation that shares the following mentioned properties can also be used.

Definition 2: (Extraction Structure Model)
Given a sentence $sentence_{ijk} \in paragraph_{ij}$, the relations are extracted with $relation_p$ as shown below:

$$\begin{aligned} &<< \\ &\quad relation; \\ &\quad argrole_1 : entity_1; word_1 : words; \\ &\quad argrole_2 : entity_2; word_2 : words; \\ &\quad argrole_3 : entity_3; word_3 : words; \\ &>> \end{aligned}$$

The relations mentioned are extracted from the parsing result of the HPSG framework. The relation tuple is resided in the SEM content of the feature structure of the parsed sentence.

Example Consider the entry in Figure 2. For the first paragraph, "Reading is a way of getting information from something that is written. Reading involves recognising ...", some of the resulting static relations extracted are:

$<< is; argrole_1 : w_{reading}; argrole_2 : t_1, argrole_3 : t_2 >>$
$t_1 :<< get; argrole_1 : w_{information} >>$
$t_2 :<< is; argrole_1 : something;$
$argrole_2 : w_{written} >>$

$<< Recognize; argrole_1 : w_{reading}; argrole_2 : t_3 >>$
$t_3 :<< make; argrole_1 : w_{language} >>$

$<< Get; argrole_1 : w_{hearing}; argrole_2 : w_{information} >>$

$<< include; argrole_1 : t_4; argrole_2 : w_{entertainment} >>$
$t_4 :<< Gain; argrole_1 : w_{information};$
$argrole_2 : w_{reading} >>$

These extracted relations represent the local relationships between the entities and relations within the sentences. In the extracted relations, there are some parameter like t_x that represent the case in which a relation fills the role of other relations.

Induction

After extracting the static relations from sentences, the next step is to link the different entities and different relations from the extraction phase.
The most intuitive approach is to find all the relations and entities with the same name and linked them together. However, this approach does not work. Like normal texts, the texts within the encyclopedia also contain incoherent relations in which the same set of entities exists in a mutual exclusive set of relations. For example, it is common to have 3 entities named e_1, e_2, e_3 existing in $relation$ and negation of this relation. This is due to the fact that the encyclopedia texts are edited by different authors and at different instance of time and space. Thus, they may not use the completely consistent set of words and concepts to describe things. The link structure of the encyclopedia provides an invaluable evidence to resolve this issue. Instead of taking a global view of relations and entities, we take a partial view of entities and relations in which the relations between two closely linked pages are induced first.

By executing the induction algorithm as shown in Figure 3 on every entry in the encyclopedia texts, each page p_i will contain an extra set $relation$ containing consistent relations from the other entries to this entry. The step of Generation will be executed based on the extraction mechanism

Algorithm 1: (Induction Phase)

Let $relation$ = $\emptyset$

Given p_i =<$\{paragraph_{ij}\}, \{insense_n\}$>

Let $relation_{self}$ = {Relation Extracted from p_i }

For $\forall insense \in insense_n$,
  For $\forall word_{\alpha\beta\gamma\delta} \in sentence_{\alpha\beta\gamma}$
  and $sentence_{\alpha\beta\gamma} \in paragraph$
  and $paragraph \in insense$
  and $word_{\alpha\beta\gamma\delta} = p_i$
    Generate $relation_{new}$ from $sentence_{\alpha\beta\gamma}$
    If $relation_{new}$ is consistent with $x \in relation_{self}$
      $relation = relation \bigcup relation_{new}$
      set $relation_{new}.argrole_x = p_i$
    end if
  end loop
end loop

Figure 3: Algorithm description of the induction phase

as shown in the previous section. The newly generated relations will be compared against the set of relation induced for the current entry by the level coherency. This step is necessary as the texts within different entries may contain inconsistent information and relations. The filtering step is performed to minimize the level of inconsistencies within the set of relations. The appropriate argument role of the newly generated relation will be set to point to the current entry. The resulting group - $relation$, contains a coherent set of entities and relations as induced from the encyclopedia.
The level of consistency is measured by whether the newly generated relation $relation_{new}$ contradicts with any of the relations in the current entry $relation_{self}$. The current algorithm uses a simple matching method to filter relations. For example, if the current entry contains the text "Reading is mostly done from paper ..." is checked against the incoming entry containing the texts "Reading is not done from paper ...". The incoming relation is filtered out. More sensible method should be based on how different the entities in the argument role between the $relation_{new}$ and $relation_{self}$ of the same type of relation. In our current framework, we adopt his simple checking method since we are still exploring this research topic.
The result of this step will be a set of relations, $relation$, with each element containing an appropriate argument role pointing to the current entry.
Example Following the figure as shown in Figure 2 and consider the entry from "school" containing the sentences "For example: writing, reading, and calculating numbers (maths). Many schools also teach arts such as music and art."
The sentence is first translated into the relation tuples, $relation_{new}$ and then matched against the entry "reading". As the entry is consistent, it is added to the $relation$. The

appropriate argument role of $relation_{new}$ will point to the entry "reading".

Generalization

The induction phase groups different entities and relations into relationships as modeled by the $relation$ set. However, further generalizations on these extracted relations are needed.
The generalization phase, as shown in Figure 4 can be thought of deducing the cross-cutting properties in the induced entities and relations. The generalized items act as the basis for further general reading tasks.

Algorithm 2: (Generalization Phase)

From the induced group of a particular sense:
  $relation_{new}$ of p_i

$role_1 = \emptyset, role_2 = \emptyset, role_3 = \emptyset$
$rel_{set} = \emptyset$

For $\forall r \in relation_{new}$,
  if $p_i \in \{argrole_x\}$where $x \in \{1, 2, 3\}$ then
    $role_x = role_x \bigcup r$
  end if
  if $p_i \in \{r\}$ then
    if r is not duplicate then
      $rel_{set} = rel_{set} \bigcup r$
    end if
  end if
end loop

For $role_x$ where $x \in \{1, 2, 3\}$
  $val = \left(\frac{tuplenum_y}{totalnum_y}\right) + \left(\frac{tuplenum_z}{totalnum_z}\right) + \left(\frac{tuplenum_{rel}}{totalnum_{rel}}\right)$
  Define class c_x where c_x contains
  the tuple $< R, role_y, role_z, val >$
  where $x \neq y$ and $x \neq z$
end loop

For $r \in rel_{set}$
  $val = \left(\frac{tuplenum_x}{totalnum_x}\right) + \left(\frac{tuplenum_y}{totalnum_y}\right) + \left(\frac{tuplenum_z}{totalnum_z}\right)$
  Define class c_r where c_r contains
  the tuple $< role_x, role_y, role_z >$
end loop

Figure 4: Algorithm description of the Generalization Model

Categorization Function:

The role of the categorization function is to deduce how similar two entities are within the possible world of encyclopedia. After executing the generalization phase, each sense will have three new classes containing the information of how they are related to other senses. These relations can be of the three argument roles or in the relation. A value for each tuple is calculated based on the following formula:

$$val = \sum \left(\frac{tuplenum_x}{totalnum_x}\right)$$

where $tuplenum_x$ is the number of occurrence of $role_x$ within a particular sense; and
Let $totalnum_x$ is the number of occurrence of $role_x$ within the corpus.

These values measure how representative a pointing entry is with respect to p_i. A more representative pointing entry means that within the corpus, or within the structure of the world spanned by the corpus, the pointing entry and p_i are more frequently related with respect to other entries with a less value and will be used in reading more general texts.

Extension with new text documents

In encountering the more general text, in which, unlike encyclopedia, no sense linkage exists. All the information a particular algorithm has is the surface order of the words within the sentences and the paragraph information.
Using the similar strategy of the induction phase, the basic unit of processing in handling the general text is a paragraph and the notation of the paragraphs, sentences and words are defined as below, except the words now have no linkage information. Also, the sense p_i is undefined as the paragraph now contains a lot of different senses, some may be conflicting.

Algorithm 3: (Extension with new texts)

Assume $paragraph_a$ is being processed:
For $\forall sentence_{ab} \in paragraph_a$
  Extract the set of $\{relation\}$ from $sentence_{ab}$
  For $r \in relation$
    For $\forall$ known entities within r,
      Check whether r and
      the relations in the $relation$ collected in
      generalization phase are consistent.

      If relations are consistent,
      add this relation to the appropriate $relation_{set}$
    end loop

    For unknown entities,
      Form new relation $relation_{add}$ containing
        these entities.
      If $relation_{add}$ contains some entities
      existing in the corpus
        Build up a new link from this relation
        to the target entities.
      end if
    end loop

  end loop
end loop

Figure 5: Algorithm description of the Extension Phase

The extension algorithm as shown in Figure 5 tries to expand the current encyclopedia knowledge base with the

newly processed text. Theoretically, by processing more texts, the base will become larger and can cover more relations and entities in processing.

Conclusions and Future Work

In this paper, we presented a mechanism in which different entities and relations from the real world texts can be collected and related in a systematic way by exploiting the online encyclopedia. Through tracing the sense linkages and the underlying texts, the entities and relations existing in the world is gradually built up. These relationships between the induced entities and generalized relations are reflecting the way the world is structured and thus provide a ground for further reasoning and inference mechanism. Though the more general mechanism of reading raw texts is still a great challenge to be overcome, we believe a structured knowledge base of the world built up with mechanism as proposed by this paper can provide more sensible and accurate reasoning performance when performing general text reading.
We will continue the further development of the model described here. We will investigate a more complete algorithm in dealing with more general texts with the relationships of the entities and relations induced from the encyclopedia. Also, the feasibilities of utilizing the external links of the encyclopedia entries to further improve the set of entities and relations will be investigated so that an larger set of entities and relations can be grouped for further processing.

References

Bresnan, J. 2001. *Lexical-functional syntax*. Blackwell Publishers Inc.

Chen, J.; Ji, D.; Tan, C.; and Niu, Z. 2006. Semi-supervised relation eextraction with label propagation. In *Proceedings of the Human Language Technology*.

Clark, P.; Harrison, P.; Jenkins, T.; Thompson, J.; and Wojcik, R. 2006. From wordnet to a knowledge base. In *AAAI Spring Symposium*, 10–15.

Pollard, C., and Sag, I. 1994. *Head-Driven Phrase Structure Grammar*. CSLI Lecture Notes Series.

Schubert, L. 2002. Can we derive general world knowledge from texts? In *Proceedings of the Human Language Technology*, 84–87.

Smith, N., and Eisner, J. 2004. Annealing techniques for unsupervised statistical language learning. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 487–494.

Widdows, D. 2003. Unsupervised method for developing taxonomies by combining syntactic and statistical information.

Zhao, S., and Grishman, R. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.