Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet

Marine Carpuat
Grace Ngai
Pascale Fung
Kenneth W.Church

About this paper

Creates a bilingual ontology by aligning WordNet with an existing Chinese ontology HowNet

- Borrows techniques used in *information retrieval and machine translation.*
- Wants to show there exists an efficient algorithm that is capable aligning ontologies with two very different language structures
- Structural information within the ontologies
  - Not applicable to ontology that have vastly diff. structure

A Bilingual Chinese-English ontology

- Linking the American English WordNet and Simplified Chinese HowNet together by their most basic concepts
  - the WordNet synset and the HowNet Definition.
- Why picked WordNet & HowNet?
  - Structure
  - Polysemous words
  - Excellent test for the portability of the algorithm

WordNet

- Electronic lexical database
- Differentiate word senses from each other through the use of synsets.

  Ex: “address” -- {address, computer address}, {address, speech}
- Synsets are linked to other synsets through hierarchical relations. (ex: hyponyms, hypernyms)
- A total of 109,377 synsets are defined.

HowNet

- Electronic lexical database
- Mostly in Chinese with some English technical terms (ex: ASCII)
- Synsets are not explicitly defined
  - Many words often belongs to the same definitions
- 1500 basic definitions
- A total of 16,788 word concepts are composed of subsets of the definition

Want to know more?

- A detailed WordNet –HowNet Structural comparison can be found in Wong & Fong (2002)

Word Sense ambiguation problem

- Finding the correct translation for Polysemous word in Chinese and English was the biggest problem.
  - Example: “Crane”

- One can see the problem of ambiguation by :
  - Baseline Experiment:

    Step 1: Pick 2000 HowNet definitions (and associated words) at random

    Step 2: Translate each of these words to English

    Step 3: Associate each of the translated English words with one synset in WordNet.

Result of Baseline Experiment

Average no. of HowNet Entries per Definition	5.4
Average no. of WordNet Synsets per Definition	8.1

- For every definition in HowNet, there are on average 5 Chinese words with that definition
- For every definition in HowNet, there are on average 8 WordNet associated synsets.

Finer-Mapping Approach…

- **Definition Match Algorithm (Knight & Luk, 1994)**
  - Compare words with their contexts from example sentences and definition found in a dictionary.
  - Uses word contexts from a large bilingual corpus.
- **Fung & Lo 's information retrieval-like method**
  - Comparison of word contexts across languages and corpora that need not be parallel
  - Effective at extracting bilingual word trans. pairs

$$similarity\ (w_e, w_c) = \frac{\sum_{i=1}^{t} (w_{ic} \times w_{ie})}{\sqrt{\sum_{i=1}^{t} w_{ic}^2 \times \sum_{i=1}^{t} w_{ie}^2}} \times \frac{2\sum_{i=1}^{t} (w_{ic} \times w_{ie})}{\sum_{i=1}^{t} w_{ic}^2 + \sum_{i=1}^{t} w_{ie}^2}$$

where

$$w_{ic} = TF_{ic} \times IDF_i$$

$$w_{ie} = TF_{ie} \times IDF_i$$

Using Synsets for Word Sense Disambiguation

Goal of the algorithm:

The alignment of the proper translation pair to the correct word sense

- The candidate WordNet synsets are ranked according to their similarity with the Chinese HowNet definition.
- The alignment 'winner' is defined as the HIGHEST-RANKING WordNet synset.

Word Sense Alignment Method …

1. Given a HowNet definition d, first extract its associated set of Chinese words and their English translations.

2. For each word from the English translations, find all the WordNet synsets that it belongs to.

3. For each of these candidate WordNet synsets s,
   a) If s contains only a single word(|s| = 1), expand it by adding words from its direct hyperset*.
   b) Define:

$$similarity(d,s) = \frac{\sum_{w_e \in s} \sum_{w_c \in d} similarity(w_e, w_c)}{\sum_{w \in s} appears(w)}$$

where $$appears(w) = \begin{cases} 1 & \text{if } n_w > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is hyperset?

- The set of hypernyms of the current word which are included to aid in defining the meaning.

Why need it?

- The algorithm works better with synsets that contains more entries.
  - More elements in the Synsets , the greater of the value of Similarity (d,s).

Experiment…

- Bilingual data source: English-Chinese Hong Kong News Corpus which comprises of 18,500 aligned article pairs, from news doc released between 1997-2000.
  - * over 6 million words on the English side
  - * use the entire HowNet vocabulary as a lexicon.
- The word list for the context vector construction was extracted by taking the monosemous (single meaning) word from WordNet
- Throw out all the words that had more than one translation in Chinese

Overall Result

- For each HowNet definition , the highest scoring WordNet synset that was aligned to it, and the corresponding alignment score are shown.
- The reverse mapping of WordNet synsets to HowNet definitions can also demonstrate the capabilities of the method.

HowNet Definition	Top Aligned WordNet Synset(s)	Score
human|人, #occupation|职位, employee|员	{employee, worker}	0.002456
BeNot|非	{ name, identify}	0.002311
human|人, unable|庸, undesired|莠	{master, original}	0.0007193
BeRecovered|复原,StateIni=alive|活着	{revive}	0.0004365
image|图像, $carve|雕刻	{sculpture}	0.0003106
AlterForm|变形状	{top, pinch}	0.0001777
aValue|属性值,rank|等级,elementary|初	{elementary, primary}	0.0001083
AimAt|定向	{calculate, aim, direct}	8.958×10^{-5}
attribute|属性, pattern|样式, physical|物质	{form, word form}	4.859×10^{-5}
break|折断	{break}	4.624×10^{-5}
pay|付,possession=money|货币	{pay}	3.769×10^{-5}
BeGood|良态	{state}	3.739×10^{-5}
BeOpposite|对立	{confront}	1.460×10^{-5}
donate|捐, possession=money|货币	{subscription}	1.094×10^{-5}
HoldWithHand|搀扶	{pass, hand, reach, pass on, turn over, give}	4.9565×10^{-6}
AmountTo|总计, means=CauseToBe|使之是	{convert, change over}	2.557×10^{-6}
time|时间,@rest|休息,education|教育	{break, pause, interruption}	2.173×10^{-6}
Avalue|属性值,form|形状,even|匀	{even}	1.549×10^{-6}
BeBad|衰变	{die, decease, perish, go, exit, pass away, expire}	1.792×10^{-7}
AlterLocation|变空间位置	{exchange, change, interchange}	1.4333×10^{-7}

Table 2: Top Ranking Alignments of HowNet definitions to WordNet Synsets. (Words enclosed in curly braces belong to the same synset)

Final Analysis

- 1-to-1 mapping from all HowNet definitions to WordNet synsets does not exists
- The seed word (a word that can be directly translation from one lang. to the other) coverage
  - ✓Precise translation? (!! No !!)
  - ✓What about Rare Words? It creates lots of blank fields.
- Non-compositional compounds (NCC) causes problem
  - ✓Ex: floppy disk, hot dog
- Implement stemming technique
  - ✓Be able to capture the way a word is used more accurately

Conclusion and Future Work

- Does not make any assumptions about the structural alignment between both ontologies
- Expand the work on:
  - Address the concerns in the analysis section
  - Produce a full alignment from HowNet to WordNet
  - Expand the algorithm with more structural info.
  - Examine the use of the aligned ontology in application (cross-lingual information retrieval and machine translation)