Identifying Concepts Across Languages: A First Step towards a Corpus-based Approach to Automatic Ontology Alignment

Grace Ngai[†] **grace@intendi.com**

Marine Carpuat[‡] **eemarine@ust.hk**

Pascale Fung[†,‡] **pascale@ee.ust.hk**

[†]Intendi Inc.
Hong Kong

[‡]Human Language Technology Center
HKUST
Clear Water Bay, Hong Kong

1 Introduction

The growing importance of multilingual information retrieval and machine translation has made multilingual ontologies an extremely valuable resource. Since the construction of an ontology from scratch is a very expensive and time consuming undertaking, it is attractive to consider ways of automatically aligning monolingual ontologies, which already exist for many of the world's major languages.

This paper presents a first step towards the creation of a bilingual ontology through the alignment of two monolingual ontologies: the American English WordNet and the Mandarin Chinese HowNet. These two ontologies have structures which are very different from each other, as well as being constructed for two very different languages, which makes this an appropriate and challenging task for our algorithm.

2 Alignment of Ontologies

In this paper, we address the problem of automatic multilingual ontology alignment. Multilingual ontologies are very useful, but are also very time-consuming and expensive to build. For example, Euro WordNet (Vossen, 1998), a multilingual ontology for 8 European languages, involved 11 academic and commercial institutions and took 3 years to complete. Furthermore, for many of the world's major languages, monolingual ontologies already exist in some shape or form. Therefore, it is reasonable and attractive to investigate whether a multilingual ontology could be quickly and robustly constructed from monolingual resources.

Given the easy availability of bilingual dictionaries, the task might seem easy at a first blush. However, given two independently constructed ontologies, there always exists some difference in their structure that makes it difficult to perform a purely structural alignment. These differences arise from different approaches and philosophies taken during the construction of the ontology; and for ontologies in different languages, differences which stem from dissimilarities between the languages concerned.

In addition, multilingual ontology alignment also has to deal with machine translation issues. Since an ontology arranges words in a semantic hierarchy, it is possible for a word to appear in several different places in the hierarchy depending on its semantic sense. However, words and concepts in a given language do not always translate cleanly into a second language; a word often has multiple translations, and they do not always share the same meanings. In the absence of any ambiguity resolution, synonym sets in one ontology will be erroneously aligned to multiple synonym sets in the second ontology. This is a serious problem: an investigative experiment with two ontologies, the American English WordNet and the Mandarin Chinese HowNet, found that, in the absence of any word sense disambiguation, each HowNet definition (the equivalent of a synonym set from WordNet) corresponded to an average of 8.1 WordNet synonym sets.

The approach taken in this paper works upon the assumption that even though a word may have different translations that correspond to different semantic senses, it is not likely that its synonyms will have the same exact set of translations. Given a synonym set, or *synset*, in one ontology, our approach considers the average similarity between its words and words from all potential alignment candidates:

Given two ontologies o and o', a synonym set (synset) $s \in o$, and a similarity score $sim(w_i, w_j)$ between any two words:

1. For each word $w \in s$, find the synsets in o' that it appears in (in the cross-lingual case, find the synsets in o' in which the translations of w appear.)
2. For each of these candidate synsets s':
   (a) if words $w \in s$ (or their translations) ap-

pear in the direct hyperset or hyposet, add them to s'.

(b) if s' contains a single word ($|s'| = 1$), expand it by adding words from its direct hyperset.

(c) Calculate $sim(s, s')$:

$$sim(s,s') = \frac{\sum_{w \in s} \sum_{w' \in s'} sim(w,w')}{\sum_{w \in s} appears(w)}$$

where

$sim(w,w')$ as defined in Section 3

$$appears(w) = \begin{cases} 1 & TF(w,s) \geq 1 \text{ for some } s \\ 0 & \text{otherwise.} \end{cases}$$

The candidate synsets from o' are then ranked according to their similarity with s, and the synset with the largest similarity is considered to be the alignment "winner".

3 Cross-lingual Semantic Similarity

Since automatic ontology alignment involves the comparison of sets of words to each other, it is necessary to define some measure for semantic similarity. Much work has been done on this topic, but most of it has been in monolingual semantic similarity calculation. Our problem is more complicated, as a cross-lingual ontology alignment will require measuring semantic similarity of words from different languages.

The method used in this paper is an extension of work from Fung and Lo (1998). The assumption is that there is a correlation between word cooccurrence patterns that persists across languages, and the similarity between word cooccurrence patterns is indicative of the semantic similarity. To construct a representation of the cooccurrence patterns, a list of *seedwords* is compiled. The seedwords in one language is a direct translation of those in the other language. Given a bilingual corpus, a *context vector* can then be constructed for each of the words of interest, where each element in the vector is a weight corresponding to a function of the significance of a particular seedword and its cooccurrence frequency with the word of interest. This method, which was applied to the problem of automatic dictionary induction, has the advantage of being able to utilize non-parallel bilingual corpora, which is by nature much more plentiful than parallel corpora.

The most important extension that our work makes to the work of Fung et al. is the introduction of *translation groups* of words. A major issue with translation research is that, given two arbitrary languages, it is common for a word in one language to have multiple translations in the other. It is also common for a given translation of a particular word to be a translation of one of its synonyms as well.

To address this problem, this work uses *seedword groups*, m-to-n translations of sets of words, rather than 1-to-1 translations of single words. This increases the robustness of the method, since a word need not be consistently translated for its context to be accurately identified. An additional benefit is that the sparse data problem is alleviated somewhat: the increased number of seedwords increases the coverage of the corpus, which reduces the possibility that a rare word whose translation we are interested in does not occur with any of the seedwords.

Given two languages, l_1 and l_2, the algorithm proceeds as follows:

1. Define a list $\mathcal{S}_1 = \{\mathcal{S}_{10}, \mathcal{S}_{11}, ...\mathcal{S}_{1t}\}$, where each member $\mathcal{S}_{1i}$ of the list is a set of words in l_1.

2. Create a list $\mathcal{S}_2 = \{\mathcal{S}_{20}, \mathcal{S}_{21}, ...\mathcal{S}_{2t}\}$, where $\mathcal{S}_{2i}$ is a set of words in l_2 which are translations of the words from $\mathcal{S}_{1i}$.

3. For each word w of interest in l_i, create a vector $\vec{v} = \{v_0, v_1, \ldots, v_t\}$ such that:

$$v_j = \frac{\sum_{s \in \mathcal{S}_j} weight(s)}{|\mathcal{S}_j|}$$

where

$$weight(s) = (log(TF(w,s)) + 1) \times IDF(s)$$

$TF(w,s)$ = Term frequency (number of occurrences) of w in the context[1] of s

$$IDF(s) = -log\frac{n_s}{N}$$

n_s = Number of occurrences of s in the corpus

N = Maximum number of occurrences of any seedword in the corpus

4. Given a pair of words w_i and w_j, define $sim(w_i, w_j) = \cos(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \cdot \vec{v_j}}{|\vec{v_i}||\vec{v_j}|}$

[1] For this work, the context of a word is defined to be the sentence that it appears in.

4 Experiment Details

4.1 Ontologies

The ontologies selected for alignment in this work were the American English WordNet (Miller et al., 1990) version 1.7, and the Mandarin Chinese HowNet (Dong, 1988).[2]

There are two main reasons why these particular two ontologies were chosen: they represent very different languages, and were constructed with very different approaches. WordNet was constructed with what is commonly referred to as a differential theory of lexical semantics (Miller et al., 1990), which aims to *differentiate* word senses by grouping words into synonym sets (synsets), which are constructed as to allow a user to easily distinguish between different senses of a word.

HowNet, in contrast, was constructed following a *constructive* approach. At the most atomic level is a set of almost 1500 basic definitions, or sememes, such as "human", or "aValue" (attribute-value). Higher-level concepts, or definitions, are composed of subsets of these sememes, sometimes with "pointers" that express certain kinds of relations, such as "agent" or "target", and words are associated with the definition(s) that describe them. For example, the word "疤" (scar) is associated with the definition "trace|痕,#disease|疾病,#wounded|受傷".

HowNet contains a total of almost 17000 definitions. On average, each definition contained 6.5 Chinese words, with 45% of them containing only one word, and 10% of them containing more than 10 words. Since the words within a definition are composed of the same sememe combination, HowNet definitions can be considered to be the equivalent of WordNet synsets.

A detailed structural comparison between HowNet and WordNet can be found in (Wong and Fung, 2002).

4.2 Supplementary Dictionary

To supplement the English translations included in HowNet, translations were included from CEDict, an open-source Chinese-English lexicon which was downloaded from the web. The two lexicons were merged to create the final dictionary by iteratively grouping together Chinese words that shared English translations to create our m-to-n seedword translation groups.

The finalized dictionary is used to create seed word groups for building the contextual vectors. First, the mappings in which none of the Chinese or English words appear in the corpus are filtered out. Second, only the mappings in which all of the Chinese words appear in the same HowNet definition are kept. The remaining 1975 mappings, which consist of an average of 2.0 Chinese words which map to an average of 2.2 English words, are used as seed word groups.

4.3 Corpora

The bilingual corpus from which the context vectors were constructed are extracted from newspaper articles from 1987–1992 of the American English Wall Street Journal and 1988–1996 of the Mandarin Chinese People's Daily newspaper (人民日報). The articles were sentence-delimited and a greedy maximum forward match algorithm was used with a lexicon which included all word entries in HowNet to perform word segmentation on the Chinese corpus. On the English side, the same greedy maximum forward match algorithm is used in conjunction with a lexicon consisting of all word phrases found in WordNet to concatenate individual words into non-compositional compounds. To ensure that we were working on well-formed, complete sentences, sentences which were shorter than 10 Chinese words or 15 English words were filtered out. A set of sentences were then randomly picked to be included: the final corpus consisted of 15 million English words (540k sentences) and 11.6 Chinese words (390k sentences). Finally, the English half of the corpus was part-of-speech tagged with fnTBL (Ngai and Florian, 2001), the fast adaptation of Brill's transformation-based tagger (Brill, 1995).

It is important to note that the final corpus thus generated is not parallel or even comparable in nature. To our knowledge, most of the previous work which utilizes bilingual corpora have involved corpora which were at least comparable in origin or content, if not parallel. The only previous work that we are aware of which uses unrelated corpora is that of Rapp (1995), a study on word co-occurrence statistics in unrelated German and English corpora.

5 Experiments

To get a sense of the efficacy of our method, a test set of 160 HowNet definitions were randomly cho-

[2]The entries in HowNet are mainly in Chinese with a few English technical terms such as "ASCII". English translations are included for all the words and sememes.

sen as candidates for the test set.[3] The Chinese words contained within the definitions were extracted, along with the corresponding English translations. Two sets of context vectors, C_c and C_e, can then be constructed for the Chinese words in the definition and their English translations. Once these context vectors have been constructed, the similarities between the HowNet definitions and the WordNet synsets can be calculated according to the formulae in Section 2.

6 Results

To get a sense of the complexity of the problem, it is necessary to construct a reasonable baseline system from which to compare against. For a baseline, all of the synsets that directly correspond to the English translations were extracted and enumerated. Ties were broken randomly and the synset with the highest number of corresponding translations was selected as the alignment candidate.

Because there is no annotated data available for the evaluation, two judges who speak the languages involved were asked to hand-evaluate the resulting alignments, based on, firstly, the set of sememes that make up the definition, with the words that are contained in the definition only as a secondary aid. Table 1 shows the overall performance of our algorithm, and Table 2 show the highest-scoring alignment mappings.

	Correct	Incorrect	Accuracy
Similarity	106	54	66.3%
Baseline	94	66	58.8%

Table 1: Overall Performance Figures

In addition to the overall results, it is also interesting to examine the rankings of the alignment candidates for some of the more difficult HowNet definitions.

Table 3 shows an example definition and the candidate rankings. This definition includes the words "population" and "number of people", however, "number of people" was filtered out as it does not occur in WordNet as a single collocation, leaving only "population", a noun with 6 senses in WordNet, to work with. This example is a good illustration of the strength and power of the cross-lingual word similarity calculation, as the system correctly identifies the first sense of "population" — "the people who inhabit a territory or state" — as the correct semantic sense of this particular definition from the Chinese words "人口" (number of human mouths), "人數" (number of people) and "人頭" (number of human heads).

Another very good example of the algorithm's strength can be found in the rankings for the HowNet definition "TakeAway|搬動,patient=family|家" (Table 4). Again, the phrasal word translations "move house", "change one's residence", "move to a better place", etc were filtered out, leaving the single word "move", which has a total of 16 senses as a verb in WordNet 1.7. However, as the table shows, the algorithm correctly assigns the "change residence" sense of "move" to the HowNet definition, which is appropriate for the Chinese words it contains, which include "搬家" (move house), "遷居" (change one's dwelling), and "拆遷" (tear down one's house and move).

7 Analysis

Even though the final goal of our work is to construct a full mapping from HowNet to WordNet, there will be quite a number of HowNet definitions which do not have a WordNet synset equivalent. The reason is that given an arbitrary pair of languages, there will exist some words in one language which do not have a translation in the other language. In the case of English and Chinese, many of the encountered problems came from Chinese idiomatic expressions, which are common in everyday usage and are considered to be single words, but do not usually translate to a single word in English. In addition, the inherent difference in sense granularity and structure between any given two ontologies means that a full-scale mapping of synsets from one ontology to another will not usually be possible. For example, HowNet's "livestock" definition covers words which are as diverse as "cow", "cat" and "dog", while the finest-grained WordNet synset that covers all these definitions is {placental, placental mammal, eutherian, eutherian mammal}.

One of the most troublesome problems encountered in this work was in the selection of seedwords, which define set for the automatic lexicon induction. If the seedwords occur so frequently in the corpus that other words co-occur with them too easily, they will provide little useful discriminatory information to the algorithm; but if they are too rare, they will

[3] The original number of definitions chosen for the test set was higher. However, upon inspection, it was found that a number had no corresponding WordNet synset and thus cannot be aligned. The 160 are the ones which are left after the non-alignable definitions were filtered out.

HowNet definition	WordNet Synset	Similarity	Correct?
cease|停做,content=discuss|商討	adjournment dissolution ⇒ termination ending conclusion	0.416	Y
institution|機構,royal|皇,past|昔	government ⇒ system system_of_rules	0.401	Y
quantity|數量,amount|多少, &human|人	population ⇒ people	0.388	Y
place|地方,#human|人	region part ⇒ location	0.358	Y
institution|機構,police|警	police_station police_headquarters ⇒ station station_house police_office	0.349	Y
knowledge|知識,entertainment|藝	art artistic_creation artistic_production ⇒ creation creative activity	0.336	Y
knowledge|知識,#mental|精神	psychology psychological_science ⇒ science scientific_discipline	0.31	Y
agreement|條約	agreement accord ⇒ harmony accord concord concordance	0.304	N
shoot|發射,sport|體育	service serve ⇒ function work operate go run	0.287	N
bird|禽,generic|統稱	bird ⇒ vertebrate craniate	0.269	Y
attribute|屬性,distance|距離, &physical|物質	distance ⇒ region part	0.268	Y
place|地方,capital|國都, ProperName|專,(Seychelles|塞舌爾)	victoria ⇒ town	0.267	Y
suffer|遭受,content=CauseAffect|傳染	catch ⇒ surprise	0.266	N
replace|代替,content=manage|管理	corkscrew spiral ⇒ turn	0.264	N

Table 2: Top HowNet Definition to WordNet Synset alignments

quantity|數量,amount|多少,&human|人	
WordNet synset	Similarity
population ⇒ people	0.388
population ⇒ group grouping	0.336
population ⇒ colonization colonisation settlement	0.218

Table 3: Population: a group of human inhabitants, or a group of organisms?

not co-occur often enough with other words to be able to provide enough information, either. This problem can be solved, however, by a better selection of seedwords, or, more easily, simply by using a bigger corpus to alleviate the sparse data problem.

A more serious problem was introduced by the comparability of the corpora involved in the experiment. Even though both English and Chinese halves were extracted from news articles, the newspapers involved are very different in content and style: the People's Daily is a government publication, written in a very terse and brief style, and does not concern itself much with non-government affairs. The Wall Street Journal, on the other hand, caters to a much broader audience with a variety of news articles from all sources.

This creates a problem in the co-occurrence patterns of a word and its translations. The assumption that word co-occurrence patterns tend to hold across language boundaries seems to be less valid with corpora that differ too much from each other. An observation made during the experiments was some words occurred much more frequently (relative to the half of the corpus they were in) than their translated counterparts. The result of this is that their context vectors may not be as similar as desired.

The difference in the co-occurrence patterns between the two halves of the corpora are best illustrated by a dotplot (Church, 1993). The total term frequency (TF) of each seedword group is plotted against that of its translations.

Figure 1 shows the resulting dotplot. If the two halves of the corpora were exact copies of each other, the frequencies of the seedwords would be equal and the points would therefore be aligned along the $x = y$ diagonal. The further the points diverge from the diagonal, the more different the two halves of the corpus are from each other. This

TakeAway|搬動,patient=family|家	
WordNet synset	Similarity
move (Sense 4 of move — to change residence)	0.205
travel go move locomote	0.185
affect impress move strike	0.166

Table 4: Move: to change residence, to travel, or to touch?

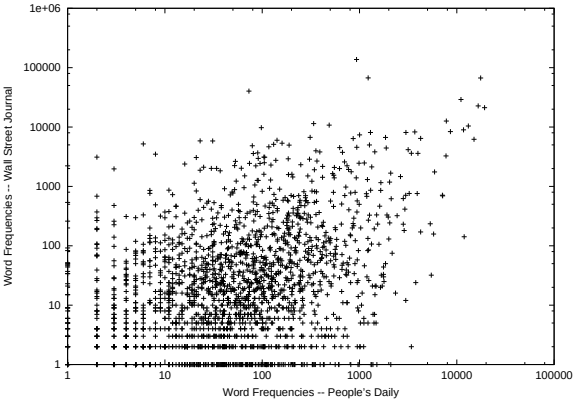

Figure 1: Seedword Group Occurrence Frequencies on People's Daily and Wall Street Journal Corpora

is quite obviously the case for this particular corpus — the overall point pattern is fan-shaped, with the diagonal only faintly discernible. This suggests that the word usage patterns of the English and Chinese halves of the corpus are quite dissimilar to each other.

It is, of course, reasonable to ask why parallel or comparable corpora had not been used in the experiments. The reason is, as noted in Section 2, that noncomparable corpora are easier to come by. In fact, the only Chinese/English corpus of comparable origin that was available to us was the parallel Hong Kong News corpus, which is about half the size. Furthermore, the word entries in HowNet were extracted from Mandarin Chinese corpora, which differs enough from the style of Chinese used in Hong Kong such that many words from HowNet do not exist in the Hong Kong News corpus. Feasibility experiments with that corpus showed that many of the seedwords either did not occur, or did not co-occur with the words of interest, which resulted in sparse context vectors with only a few non-zero co-occurrence frequencies. The result was that the similarity between many of the candidate WordNet synset-HowNet definition pairs was reduced to zero.

Despite all these problems, our method is successful at aligning some of the more difficult, single-word HowNet definitions to appropriate WordNet synsets, thus creating a partial mapping between two ontologies with very different structures from very different languages. The method is completely unsupervised and therefore cheap on resource requirement — it does not use any annotated data, and the only resource that it requires — beyond the ontologies that are to be aligned — is a bilingual machine-readable dictionary, which can usually be obtained for free or at very low cost.

8 Previous Work

The preceding sections mentioned some previous and related work that targets the same problem, or some of its subproblems. This section takes a closer look at some other related work.

There has been some interest in aligning ontologies. Dorr et al. (2000) and Palmer and Wu (1995) focused on HowNet verbs and used thematic-role information to align them to verbs in an existing classification of English verbs called EVCA (Levin, 1993). Asanoma (2001) used structural link information to align nouns from WordNet to an existing Japanese ontology called Goi-Taikei via the Japanese WordNet, which was constructed by manual translation of a subset of WordNet nouns.

There has also been a lot of work involving bilingual corpora, including the IBM Candide project (Brown et al., 1990), which used statistical data to align words in sentence pairs from parallel corpora in an unsupervised fashion through the EM algorithm; Church (1993) used character frequencies to align words in a parallel corpus; Smadja et al. (1996) used cooccurrence functions to extract phrasal collocations for translation, and Melamed (1997) identified non-compositional compounds by comparing the objective functions of a translation model with and without NCCs.

The calculation of word semantic similarity scores is also a problem that has attracted a lot of interest. The numerous notable approaches can usually be divided into those which utilize the hierarchical information from an ontology, such as

Resnik (1995) and Agirre and Martinez (2002); and those which simply use word distribution information from a large corpus, such as Lin (1998) and Lee (1999).

9 Conclusion

This paper represents a first step towards a corpus-based approach for cross-lingual identification of word concepts and alignment of ontologies. The method borrows from techniques used in machine translation and information retrieval, and does not make any assumptions about the structure of the ontology, or use any but the most basic structural information. Therefore it is capable of performing alignments across ontologies of vastly different structure. In addition, our method does not require the use of parallel or even comparable corpora, making the task of data acquisition far easier.

We demonstrate the effectiveness of our method by performing a partial mapping of HowNet and WordNet, two very different ontologies from very different languages. Our method is successful at mapping a number of HowNet definitions — including some fairly difficult ones — to the correct WordNet synsets.

10 Acknowledgements

The authors would like to thank researchers at Intendi Inc. — Ping-Wai Wong for help on HowNet construction and structure, Chi-Shun Cheung and Chi-Yuen Ma for assistance in resource preparation, as well as the three anonymous reviewers for their useful comments and suggestions.

References

E. Agirre and D. Martinez. 2002. Integrating selectional preferences in WordNet. In *Proceedings of the first International WordNet Conference*, Mysore, India.

H. Asanoma. 2001. Alignment of ontologies: Wordnet and goi-taikei. In *Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, PA.

E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565.

P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P. Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

K. Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual ACL Conference*, pages 1–8, Columbus, Ohio.

Z. Dong. 1988. Knowledge description: What, how and who? In *Proceedings of International Symposium on Electronic Dictionary*, Tokyo, Japan.

B. Dorr, G.A. Levow, and D. Lin. 2000. Large-scale construction of a Chinese-English semantic hierarchy. Technical report, University of Maryland, College Park.

P. Fung and Y.Y. Lo. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual ACL Conference*, pages 414–420, Montreal, Canada.

L. Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Conference of the Association for Computational Linguistics*, pages 25–32, College Park, MD.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada, August.

I.D. Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of EMNLP-1997*, Providence, RI.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

G. Ngai and R. Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of the 39th Annual ACL Conference*, Pittsburgh, PA.

M. Palmer and Z. Wu. 1995. Verb semantics for English-Chinese translation. *Machine Translation*, 10(1-2):59–92.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual ACL Conference*, pages 320–322.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

F. Smadja, K.R. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):3.

P. Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Press.

P.W. Wong and P. Fung. 2002. Nouns in wordnet and hownet: An analysis and comparison of semantic relations. In *Proceedings of the 1st International Conference on Global Wordnet*, Mysore, India.