# Multilingual Wordnet sense Ranking using nearest context

**E Umamaheswari Vasanthakumar** and **Francis Bond**
School of Humanities
Nanyang Technological University
Singapore
umavasanth28@gmail.com,bond@ieee.org

## Abstract

In this paper, we combine methods to estimate sense rankings from raw text with recent work on word embeddings to provide sense ranking estimates for the entries in the Open Multilingual WordNet (OMW). The existing Word2Vec pre-trained models from Polygot2 are only built for single word entries, we, therefore, re-train them with multiword expressions from the wordnets, so that multiword expressions can also be ranked. Thus this trained model gives embeddings for both single words and multiwords. The resulting lexicon gives a WSD baseline for five languages. The results are evaluated for Semcor sense corpora for 5 languages using Word2Vec and Glove models. The Glove model achieves an average accuracy of 0.47 and Word2Vec achieves 0.31 for languages such as English, Italian, Indonesian, Chinese and Japanese. The experimentation on OMW sense ranking proves that the rank correlation is generally similar to the human ranking. Hence distributional semantics can aid in Wordnet Sense Ranking.

## 1 Introduction

Most of the existing Word-net sense rankings (Navigli, 2009) use document level statistics to find the prominent sense of the given word. McCarthy and Carroll (2003) showed that predominate senses could be learned from a sufficiently large corpus, and this work has since been extended by various researchers. Words that appear nearest to the given word convey the context/meaning of a word (Lim, 2014; Liu et al., 2015; Pocostales, 2016; Rong, 2014; Long et al., 2016), and this can be used to estimate the most frequently used senses. This proposed work uses nearest context words to predict the senses and computes the frequency of occurrence of these senses within the corpus. Since most of the existing WSD systems utilize the Most Frequent Sense (MFS) as a baseline, it is important to rank the Wordnet senses in a meaningful way.

Two well-known software packages used to train word embeddings, are Word2Vec (Mikolov et al., 2013; Rong, 2014) and Glove model (Pennington et al., 2014). Polyglot (Al-Rfou et al., 2013) is a natural language pipeline that supports many NLP based tasks such as tokenization, Language detection, Named Entity Recognition, Part of Speech Tagging, Sentiment Analysis, Word Embeddings, Morphological analysis and Transliteration for many languages. This work utilizes their Word embeddings. Existing polyglot word embeddings (Al-Rfou et al., 2013) support 137 languages. We have planned to use the word embeddings for the 35 hand-built wordnets currently in OMW (Ruci, 2008; Elkateb et al., 2006; Borin et al., 2013; Pedersen et al., 2009; Simov and Osenova, 2010; Gonzalez-Agirre et al., 2012; Pociello et al., 2011; Wang and Bond, 2013; Huang et al., 2010; Pedersen et al., 2009; Fellbaum, 1998; Stamou et al., 2004; Sagot and Fišer, 2008; Ordan and Wintner, 2007; Mohamed Noor et al., 2011; Isahara et al., 2008; Montazery and Faili, 2010; Lindén and Carlson., 2010; Garabík and Pileckytė, 2013; Vossen and Postma, 2014; Piasecki et al., 2009; de Paiva et al., 2012; Tufiş et al., 2008; Darja et al., 2012; Borin et al., 2013; Thoongsup et al., 2009; Pianta et al., 2002; Oliver et al., 2015; Raffaelli et al., 2008; Toral et al., 2010).

We use corpus based frequencies for five of these languages (English, Chinese, Japanese, Italian and Indonesian) from the NTU Multilingual Corpus (NTU-MC: Tan and Bond, 2013) and use them to evaluate the learned sense rankings. Our major contribution is training and testing on large numbers of multiword expressions, which are often neglected in the word embedding literature. We identify the multi-word expressions found in the hand-built lexicons and train our own model for them using Word2Vec (Mikolov et al., 2013; Rong, 2014) and Glove (Pennington et al., 2014).

This paper is structured as follows. Section 2 discusses the related work in Word embedding and its application in WordNet Synset Ranking. Section 3 describes the data, methods, and Section 4 discusses the evaluation of results obtained from word embedding and its effect in WordNet Sense Ranking. Finally, Section 5 concludes with the findings and future plans to improve the results.

## 2 Related Work

Word embedding techniques have been popular in recent years in Word Sense Disambiguation *(WSD)* research. Similar to this proposed work, (Bhingardive et al., 2015b) computes word embeddings with the help of pretrained Word2Vec(Mikolov et al., 2013; Rong, 2014) and matches with the sense embeddings obtained from the Wordnet features. They have attempted Wordnet sense ranking for Hindi and English. Since the Word2Vec (Mikolov et al., 2013; Rong, 2014) model is based on the words frequency of occurrence in the corpus, finding the nearest context words that occur infrequently in the corpus is difficult.

Panchenko (2016) compares sense embeddings of AdaGram (Bartunov et al., 2015) with BabelNet (Navigli and Ponzetto, 2010) synsets and proved that sense embeddings can be retrieved by automatically learned sense vectors. Sense embeddings for a given target word are identified by finding the similarity between the AdaGram Word embeddings list with the BabelNet Synsets words list. Rothe and Schütze (2015) proposed an approach that takes word embeddings as input and produces synset, lexeme embeddings without retraining them. They used WordNet lexical resource to improve word embeddings.

Arora et al. (2016) showed that word vectors can capture polysemy and word vectors can be thought of as linear superpositions of each sense vector. They have attempted discourse analysis to find the cluster of sense vectors.

Although the basic idea of word embeddings is not tied to any one languages, the preprocessing steps are language specific. Kang et al. (2016) presented a cross-lingual word embedding for English and Chinese Word Sense Disambiguation (WSD). They have experimented with the performance of WSD using different word embeddings such as Word2Vec (Mikolov et al., 2013; Rong, 2014) and Glove model. Bhingardive et al. (2015a) compared word embeddings obtained from the Word2Vec (Mikolov et al., 2013; Rong, 2014) model and the sense embedding obtained from the WordNet for English and Hindi languages and restricted to Nouns. They used various WordNet features similar to this proposed work to find the predominant sense. Their approach outperforms SemCor baseline for words with the frequency below five.

In this research context words are identified with the help of Polyglot(Al-Rfou et al., 2013) word embeddings.

## 3 Methodology

In this work, we use word embeddings to find the nearest context of a given word and compare it with the senses obtained from the OMW to find the most frequently used senses. Our aim is to rank the senses obtained from the OMW with the help of the context words and their frequency of occurrence. Initially, we use the pretrained polyglot word embedding model (Al-Rfou et al., 2013) to retrieve the nearest context words and found multiwords are unidentified. Hence in this work, we have trained our own model similar to polyglot for both single and multi-words. Our aim is to train this model for all 35 languages supported by OMW, for this paper we present only the results for the five languages for which we have evaluation data: English, Chinese, Japanese, Italian and Indonesian.

### 3.1 Corpus Cleaning and preprocessing

We exploit the openly available Polyglot wiki dump corpus (Al-Rfou et al., 2013) for English, Chinese, Japanese, Italian and Indonesian. We chose this as it contains various domains and languages. Before training our own model, the corpus texts are preprocessed by removing symbols, numbers and shortest text. Stop words have been removed with the help of the NLTK toolkit (Bird et al., 2009). However, NLTK does not support stop-words for all languages. Hence we have included stop words of Chinese, Japanese, Indonesian, Italian from publicly available online utilities to NLTK toolkit. For English, Indonesian and Italian we have lemmatized each word of the cleaned text to find their base form. Chinese does not inflect, and Japanese inflections are normally split off by the tokenizer. Hence we have used Mecab to tokenize/lemmatize Japanese texts. After preprocessing the text, each sentence of the corpus is tokenized into single and multiple terms. In order to identify the multiwords from the corpus, we have used the existing Wordnet MWE lexicon *(MWEs)*. The terms of each sentence are matched with the existing wordnet *MWEs* lexicon and if an MWE is found it is rewritten to a single token, with spaces replaced by an underbar "_" symbol. The preprocessed MWE tagged texts are given as input to train our own model. So, for example, a sentence like *I looked five words up* will be preprocessed to `I look_up word`.

### 3.2 Training Model

Word embeddings for the above five languages have been trained using the Polyglot2 (Al-Rfou et al., 2013) package and *Global Vectors for Word Representation Glove Model* (Pennington et al., 2014). Polyglot2 is a software package that enables building your own language models. It learns the distributed representations of words/word embeddings for the given corpus. GLOVE is another unsupervised learning algorithm used for obtaining vector representations of words. Training is performed by considering global word-word co-occurrence statistics from a corpus and results with the linear substructures of the word vector space. We can build our own word embeddings with the help of Polyglot2 and Glove models.

### 3.3 Predominant Sense Scoring

To find the predominant senses for the given word $w$, the senses obtained from the OMW are represented as $S_w = S_1, S_2, \ldots, S_n$. The neighbouring context obtained from Polyglot2 or Glove is represented as $S_w^N(w, d)$ where $N$ represents the number of neighbouring contexts from word embedding obtained for the senses $S_w$ that can vary from 1 to $N$, and $d$ represent the distance score between the $S_w$ and $S_w^N$. $P_s(S_w)$ represents the predominant score of $S_w$ based on the WordNet synset similarity.

$$P_s(S_w) = log(sum(S_w^N(w,d)) + M_W^T/T^N W_e) + [H_s(M)/T^N W_e]) \quad (1)$$

$M_W^T$ - represents the number of matching terms between the $OMW$ synset definitions and example sentences with respect to polyglot word embeddings.
$T^N W_e$ - represents the number of word embeddings obtained from Polyglot2.
After computing the predominant score $P_s$ $(S_w)$ for each word-net entries the semantic similarity between the word embedding with the OMW ontology hierarchy is measured. $H_s(M)$ represents the number of concepts such as Hypernyms and Hyponyms of WordNet Ontology that match with the number of terms obtained in the polyglot word embeddings. The intuition behind is that the words in the word embedding will have similar words that can appear in WordNet hierarchy. For example, the word *party* may refer to a person, organization or an occasion. If it refers to a *person*, the hypernyms are *person* and the hyponyms are *assignee, assignor, contractor, intervenor.* Similarly for organization the hypernyms is *set* and hyponyms are *fatigue_party, landing_party, party_to_the_action, rescue_party, earch_party, stretcher_party, war_party* and for considering *occasion* as sense the hypernyms are *affair* and hyponyms are *bash, birthday party, bunfight, ceilidh, cocktail_party, dance, fete, house_party, jolly, tea_party, whist_drive.*

When we give *Person* as Input to Polyglot2(Al-Rfou et al., 2013), we will get the following word embeddings. *person-0.575121*, *contractor-0.628679*, *team-0.619203*, *division-0.682174*, *unit-0.700489*, *government-0.62491*, *strategy-0.725378*, *event-0.692839*, *camp-0.689145 program-0.688767*. The terms such as *person* and *contractor* matched with the Wordnet hypernyms and hyponyms. Thus *person* sense is the most predominantly used when compared to *organization* and *event* senses since it shares the semantics with WordNet hierarchy. Similarly, we can match with other features of WordNet senses to infer which sense is important.

## 4 Results and Evaluation

In this section, the word embedding models such as (Glove: Al-Rfou et al., 2013) and (Word2Vec: Pennington et al., 2014) have been evaluated on two different tasks such as word-sense ranking of Wordnet and query expansion for clinical texts, then we present some examples of word embeddings for intuitive comprehension. The word sense ranking and trained word embeddings have been tested for 5 languages English, Chinese, Japanese, Indonesian and Italian languages of Semcor dataset for the words with more than one sense. The Polyglot2 word embedding model have been trained with the Context Window Size as 14, Initial learning rate as 0.025, Hidden Layer size as 32 and minimum word count as 2 (Al-Rfou et al., 2013). Glove word embedding model has been trained with the minimum word count as 2, Vector size as 100, Maximum Iteration as 100 and Context Window size as 14 (Al-Rfou et al., 2013).
We use two metrics to measure the efficiency of the baseline and the proposed word embedding model.

- Accuracy - The fraction of relevant word embeddings among the top 10 word embeddings are measured based on the human-relevant judgment.

- Rank Biased Overlap (RBO) - The rank correlation metrics that measures similarity and dissimilarity between two ranked list.

### 4.1 Baseline

We have taken two baseline approaches. One based on the corpus frequency based approach and the other based on the Topic model distribution score (LexSemTm). Corpus frequency-based approach ranks the synset based on the frequency of occurrence of the lemma across the corpus whereas the LexSemTm used an unsupervised sense distribution learning method (LexSemTm) (Bennett et al., 2016), that utilizes *HDP-WSI* based sense learning (Lau et al., 2014). In Bennett et al. (2016), the sense distribution of words for each sense is obtained by estimating the maximum likelihood of terms with the topics.
Both the baseline approaches used SemCor Dataset. Here the SemCor Dataset is separated into groups of lemmas with frequency *1-3(Group I)*, *4-8(Group II)*, *9-20(Group III)* and greater than 21(Group IV) as described by Bennett et al. (2016). In each group, the sense distribution for each lemma is obtained from LexSemTm and the senses are ranked in descending order based on the sense distribution score and similarly for corpus frequency based method the senses are ranked based on the frequency of lemma. Then these results are compared with the proposed work.

### 4.2 Analysis on Word embedding

Evaluation was carried out on English, Japanese, Chinese, Indonesian and Italian word embedding using Polyglot2 (Word2Vec) and Glove. We found that the *Glove* model gave a better result when compared to the *Polyglot2(Word2Vec)* model. However, existing Word2Vec model Polyglot2[1] can capture the single terms well and to a very lesser degree the Multi-words are handled. In order to test this across domains, we have taken 5,611 unique terms from a clinical corpus and found that existing pre-trained model handles 1,500 terms semantically correct and the remaining 4,111 terms are not handled. The reason is pre-trained polyglot2 Word2Vec model is trained on wiki corpus and unable to scale up to the specific domain. Moreover, it is not trained for Multi-words. Some samples of semantic-based word embedding obtained

[1] `https://sites.google.com/site/rmyeid/projects/polyglot`

from the existing model in each language (Polyglot2) are listed below:
List of semantic-based word embedding obtained in each language for *Location* as query term are listed below:

- *Indonesian*
  - *lokasi(**location**)*
    *:Peta, persimpangan, pelabuhan, fondasi, celah, ruangan, wilayah, potensi, batas, otoritas-(**Map**, **intersection**, **harbor**, **foundation**, gap, room, **territory**, potential, limit, authority)*
- *Italian*:
  - *luogo(location) - Teatro, motivo, periodo, servizio, passato, punto, campo, caso, segno, paese- (Theater, pattern, period, service, past, point, **field**, case, sign, **country**)*
- *English*:
  - *Location-**site**, **map**, structure, **area**, direction, building, **locality**, settlement, line, Bridge*
- *Japanese*:
  - ロケーション *(Location)*
    *:* クルージング, デモンストレーション, 個室, バナー, ガレージ, 買い物, バルコニ, ウォーキング, ナビゲーション *-(Cruising, demonstration, **private room, banner, garage**, shopping,**balcony**, walking, navigation)*
- *Chinese*:
  - 位置 *(Location)*
    *:* 方向, 形式, 功能, 部分, 大小, 排列, 材料, 以上, 原本, 描述- *(**Direction**, Form, Feature, Section, Size, Arrangement, Material, Above, Original, Description)*

Since this proposed work has been trained for both single and multi-word expressions, we have specifically analyzed the embeddings for multi-words and the resultant samples are shown below.
*Sample List of Multi-words and Nearest Context Word*:

- *Query−English*:
  deficit_hyperactivity_disorder:
  - *attention, memory,deficit_hyperactivity_disorder, adhd,rigidly, proliferative, splinted, treat_attention, allergic_rhinitis, special*
- *Query−Japanese*:
  プリンス _ オヴ _ ウェールズ (Prince of Wales):
  - トレハラーゼ, ろかく, レゼルヴ, フリーア,
    グローヴス, レインボーカップファイナル,*mishnaic,traininfomation,* カタリココ
  - *(Trehalase, fighting, reserve, free, Groves, Rainbow Cup Final, mishnaic,traininfomation,Catalina Coco)*
- *Query−Chinese*:
  足球 _ 运动员 (soccer player):
  - 大 _ 祭台, 阅览,
    鐺, 諫, 分内事, 大捷, 新交, 縯, 井底
  - *(Large altar, learning clang,remonstrance, sub-ministry, victory, new cross, play, bottom*
- *Query−Indonesian*:
  erosi_pantai(erosion):
  - *: Mikrokimerisme, gerunggang, membuat_bangkrut, mikkeli, lille, superintendent, thur, cibinuang, operasi_boolean*
  - *(Microcimerism, rider, bankruptcy, mikkeli, lille, superintendent, thur, cibinuang, boolean operation)*
- *Query−Italian*:
  seconda_guerra_mondiale (Second World War):
  - *tisiddu,smetlivyi, pelligra, mortificava, skavronskij, tureaud, preprocessing, telemolise, quetzalctl*
  - *(Mixed with other language text)*

*Results of semantic based word embedding obtained for each language of Glove are listed below*:

- *Seconda_guerra_mondiale(Second Word War)(Italian):*
  - *prima_guerra_mondiale,scoppio, guerra,conflitto, dopoguerra,militare,bellico,militari, guerra_mondiale,sovietica (WWI, outbreak, war, conflict, war, military, war, Word war, military)*
- *jus_lemon(lemon juice)(Indonesian):*
  - *Memberikan_tenaga, Mengasamkan, operated, menguapkan, boya,memfermentasi,effektif, recoil, mwh, meluapkan.(provide power, acidity, ooperated, Evaporate, boya, ferment, effective, recoil, mwh,vent)*
- *Chinese:* 参考_资料 *(Reference_Information):*
  - 注释, 脚注, 参考, 辺, 资料 _ 来源, 内部 _ 网络, 注解, 服务 _ 设施, 参见, 出处 *(Reference _ information, Annotations, Footnote, reference, Side, Information source, Internal _ network, annotation, Service Facilities, See also, Source)*
- *English:Treadmill_test:*
  - *Stress_test, exercise, physiology,suggestion,participate, vigorous,walking,prescription,intensity*

English, Chinese and Italian word embeddings gave better results; whereas for Indonesian documents, the results are often mixed with other language texts, even though we are able to get meaningful word embeddings. We also found that the Japanese text corpus is tagged with minimal multi-word expressions and noisy. The reason is Japanese text has different writing styles that degrade the accuracy of MWE tagging because the MWE lexicon basically includes the standard scripts. Hence we need to fine tune the MWE tagging by properly filtering the character-level, word level non-standard noisy text.

| **Accuracy(Word2Vec)** | **Accuracy(Glove)** |
|---|---|
| 0.35 | 0.67 |

Table 1: Accuracy of Word embedding score for medical text(English)

The overall accuracy of the Glove model is 0.47 and Word2Vec is 0.31. Since existing polyglot model (Al-Rfou et al., 2013) handles single terms well and the trained glove model (Pennington et al., 2014) handle most of the terms meaningfully, we have planned to merge both the models to handle single and multi-terms word embeddings.

### 4.3 Scalability

In order to check, the scalability of these models in different domains, We have tested with Singapore Clinical Practical Guidelines documents of Dental, Medical, Nursing, and Pharmacy of 72 documents, available from *Ministry of Health*, Singapore (2016).[2] There are *124.2 MB* in all. The results are shown in Table 1. Again the accuracy of Glove model[3] is better when compared to the Word2Vec Polyglot learned model because Glove model computes co-occurrence statistics across the corpus whereas Word2Vec computes co-occurrence statistics within the context window size. The word embedding results also depend on the context window size and minimum frequency count. If we increase both the context window size and minimum frequency count to a certain extent, we can achieve semantically relevant word embeddings. However, the recall will be low. In order to find the optimum value to maintain precision and recall, we need to run the test with different values for few test samples. The quality and size of the corpus may also impact the results. Since clinical text contains only domain-specific terms which are unambiguous, we are able to achieve meaningful results. Whereas We found difficulty in Wikipedia dump corpus(5 languages) because it contains a lot of noisy

[2]They are online at `https://www.moh.gov.sg/content/moh_web/healthprofessionalsportal/doctors/guidelines/cpg_medical.html`.

[3]`https://nlp.stanford.edu/projects/glove/`

data. Our purpose of this work is to check, how far this distributional semantics can help in Word Sense Ranking and Clinical Information Retrieval.

Another validation on PubMed corpus have also been taken to check the scalabilty of this work. BioASQ[4] releases Word2Vec model for PubMed Abstracts of size 3.5GB (uncompressed). Their PubMed word2vec corpus consists of 10,876,004 English abstracts of biomedical articles that are publically available. We have taken a sample of PubMed corpus with 1.3 GB of data for training with our model and achieved average precision for multiword expressions as 0.55 and for single terms 0.72.

### 4.4 Quality of Ranking

To evaluate the quality of rankings produced by this method, we have compared the human/authors judgment rank *(Approach 1) A1* with three approaches such as Word embedding *(Approach 2) A2*, Corpus frequency ranking *(Approach 3) A3* and LexSemTm approach *(Approach 4) A4*. There are basically two well-defined algorithms such as *Spearman's* and *Kendall's tau* (Kumar and Vassilvitskii, 2010) rank correlation have been used to find the statistical difference in ranking. DCG *(Discounted Cumulative Gain)* (Harman, 2011) measures both relevance and ranking, whereas rank correlation helps to find statistically significant difference in order. *Webber et al (2010)* (Webber et al., 2010) proposed a method to compare ranking quality of two methods and addressed the top-relatedness issue. Since this proposed work needs to consider the concordance and discordance of ranked results based on position, We have used this measure to find the correlation between the two ranked lists. The correlation score is measured with *Approach 1* to *Approach 2*, *Approach 3* and *Approach 4* for the Semcor dataset. The statistics of test data is shown in Table 5. For example, when we give *"gleam"* as query, the resulted ranking of A1, A2, A3 are shown in Table 4, Table 2, Table 3, respectively. The rank overlapping between Approach 1 to Approach 2 and Approach 3 are calculated. Here in this example, the baseline (Corpus frequency) ranking (Approach 3) is dissimilar in all positions except the third position, whereas with human judgment (Approach 1) only the 3rd synset is moved to the last position and the remaining ranking is similar to the proposed approach (Approach 2). Hence the Rank correlation for Approach 3 to Approach 1 is *0.52* and Approach 2 to Approach 1 is *0.88*. Thus the rank quality depends on how much it is similar to the human judgment.

The results are shown in table 6. Table 7 shows the comparison of the Rank overlapping value of *A1-A2*, *A1-A3* and *A1-A4*. We found that the average correlation between A1 to A2 is greater than A1 to A3 and A1 to A4. This result provides an additional validation of our model as it demonstrates that the sense ranking can capture the sense preferred by a human. Hence the word embedding score definitely aid in wordnet sense ranking. When we analyze the rare sense words with frequency 1-3 and 4-8, the word embedding and Wordnet feature influence the results by providing most relevant result on the first hit. We have

[4]http://bioasq.lip6.fr/tools/BioASQword2vec/

| **Synsets (gleam)** |
|---|
| be shiny, as if wet |
| shine brightly, like a star or a light |
| appear briefly |
| an appearance of reflected light |
| a flash of light (especially reflected light) |

Table 2: Ranking result of Approach 2 (Proposed)

| **Synsets (gleam)** |
|---|
| ***a flash of light (especially reflected light)*** |
| ***be shiny, as if wet*** |
| appear briefly |
| ***shine brightly, like a star or a light*** |
| ***an appearance of reflected light*** |

Table 3: Ranking result of Approach 3 (Baseline - Corpus Frequency)

| **Synsets (gleam)** |
|---|
| be shiny, as if wet |
| shine brightly, like a star or a light |
| an appearance of reflected light |
| a flash of light (especially reflected light) |
| ***appear briefly*** |

Table 4: Ranking result of Approach 1 (Human)

| Languages | Lemma Count (MWs) | Lemma Count (Single words) |
|---|---|---|
| English | 2,361 | 8,187 |
| Chinese | 2,067 | 12,341 |
| Japanese | 473 | 5,289 |
| Italian | 262 | 9,606 |
| Indonesian | 1,134 | 5,178 |

Table 5: Statistics of Test data

| Languages | A1 to A3 | A1 to A2 |
|---|---|---|
| English | 0.55 | 0.75 |
| Chinese | 0.62 | 0.68 |
| Japanese | 0.64 | 0.69 |
| Italian | 0.61 | 0.67 |
| Indonesian | 0.44 | 0.56 |

Table 6: Average Rank correlation analysis between *A1 to A3* and *A1 to A2*

| Groups | Freq | Lemma Count | A1 to A2 | A1 to A3 | A1 to A4 |
|---|---|---|---|---|---|
| I | 1–3 | 1896 | 0.73 | 0.50 | 0.57 |
| II | 4–8 | 567 | 0.82 | 0.49 | 0.48 |
| III | 9–20 | 327 | 0.77 | 0.46 | 0.47 |
| IV | >20 | 124 | 0.87 | 0.49 | 0.48 |

Table 7: Average Rank correlation analysis

| Language | Lemma | First Hit Results |
|---|---|---|
| English | contact | a channel for communication between groups |
| English | intrusion | any entry into an area not previously occupied |
| English | celebration | a joyful occasion for special festivities to mark some happy event |
| English | no more | referring to the degree to which a certain quality is present |
| English | write up | a short account of the news |
| Japanese | 名人 (expert) | a person with special knowledge or ability who performs skillfully |
| Japanese | 召集 (convene) | a group gathered in response to a summons |
| Japanese | ビル (building) | a structure that has a roof and walls and stands more or less permanently in one place |
| Chinese | 适应 (adopt) | adapt or conform oneself to new or different conditions |
| Chinese | 加入 (join) | a process of increasing by addition (as to a collection or group) |
| Chinese | 修复 (repair) | restore by replacing a part or putting together what is torn or broken |
| Italian | detenzione(custody) | the state of being imprisoned |
| Italian | piuma(feather) | the light horny waterproof structure forming the external covering of birds |
| Italian | esaminare(examine) | look at carefully; study mentally |
| Indonesian | kehidupan(life) | the period between birth and the present time |
| Indonesian | barang(goods) | goods carried by a large vehicle |
| Indonesian | hanya(alone) | without any others being included or involved |

Table 8: First Hit Analysis Results

observed that the first hit obtained from each synset ranking found most appropriate when compared to LexSemTm (A4) and OMW Corpus frequency ranking (A3). A sample list of terms and the results of the first hit have been shown in table 8.

## 5 Conclusion

OMW has over 150 languages with word-nets built automatically, ranging from major languages like German or Korean for which there are no free word-nets, to smaller languages such as Volapuk. For all languages for which Polyglot has data (which is most of them) we will learn rankings and incorporate them into OMW, so that the lexicon is maximally useful for speakers of as many languages as possible. In future, we planned to extend this work to identifying missing senses by comparing the trained model over the sense-annotated corpus with the existing pre-trained models like polyglot. Since the Glove model is based on co-occurrence context, it gave better results even for a tiny corpus, hence we have planned to extend our model to sentence embedding using Glove model for finding nearest context sentences for a given synset example sentence to further improve our wordnet ranking.

**Acknowledgments**

This research was supported by the MOE Tier 1 grant *Semi-Automatic Implementation of Clinical Practice Guidelines in Singapore Hospitals* (RG25/13).

## References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192. Association for Computational Linguistics, Sofia, Bulgaria. URL `http://www.aclweb.org/anthology/W13-3520`.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. Breaking sticks and ambiguities with adaptive skip-gram. *arXiv preprint arXiv:1502.07257*, pages 47–54.

Andrew Bennett, Timothy Baldwin, Jey Han Lau, Diana McCarthy, and Francis Bond. 2016. Lexsemtm: a semantic dataset based on all−words unsupervised sense distribution learning. In *ACL (1)*. The Association for Computer Linguistics.

Sudha Bhingardive, Dhirendra Singh, Rudra Murthy, Hanumant Redkar, and Pushpak Bhattacharyya. 2015a. Unsupervised most frequent sense detection using word embeddings. In *DENVER*. Citeseer.

Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Harichandra Redkar, and Pushpak Bhattacharyya. 2015b. Unsupervised most frequent sense detection using word embeddings. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1238–1243.

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (`www.nltk.org/book`).

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211. URL `dx.doi.org/10.1007/s10579-013-9233-4`.

Fišer Darja, Jernej Novak, and Tomaž. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.

Valéria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: an open Brazilian Wordnet for reasoning. EMAp technical report, Escola de Matemática Aplicada, FGV, Brazil.

Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a wordnet for Arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to lithuanian wordnet. In Katarína Gajdošová and Adriána Žáková, editors, *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80. Lüdenscheid: RAM-Verlag. `http://korpus.juls.savba.sk/attachments/publications/lithuanian_wordnet_2013.pdf`.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Donna Harman. 2011. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure.

*Journal of Chinese Information Processing*, 24(2):14–23. (in Chinese).

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Hong Jin Kang, Tao Chen, Muthu Kumar Chandrasekaran, and Min-Yen Kan. 2016. A comparison of word embeddings for english and cross-lingual chinese word sense disambiguation. *arXiv preprint arXiv:1611.02956*.

Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 571–580. ACM, New York, NY, USA. URL `\url{http://doi.acm.org/10.1145/1772690.1772749}`.

Jey Han Lau, Paul Cook, Diana Mccarthy, Ana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.

Aiden Si Hong Lim. 2014. *Acquiring Predominant Word Senses in Multiple Languages*. Ph.D. thesis, School of Humanities and Social Sciences, Nanyang Technological University.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1501–1511.

Teng Long, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2016. Leveraging lexical resources for learning entity embeddings in multi-relational data. *arXiv preprint arXiv:1605.05416*.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267. Singapore.

Mortaza Montazery and Heshaam Faili. 2010. Automatic Persian wordnet construction. In *23rd International conference on computational linguistics*, pages 846–850.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Antoni Oliver, K. Šojat, and M. Srebačić. 2015. Automatic expansion of Croatian wordnet. In *Proceedings of the 29th CALS international conference "Applied Linguistic Research and Methodology"*. Zadar (Croatia).

Noam Ordan and Shuly Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Alexander Panchenko. 2016. Best of both worlds: Making word sense embeddings in-

terpretable. In *the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. URL `\url{http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf}`, (ISBN 978-83-7493-476-3).

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.

Joel Pocostales. 2016. Nuig-unlp at semeval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. *Proceedings of SemEval*, pages 1298–1302.

Ida Raffaelli, Božo Bekavac, Željko; Agić, and Marko Tadić. 2008. Building Croatian wordnet. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christianne Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference 2008*, pages 349–359. Szeged.

Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend:extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.

Ervin Ruci. 2008. On the current state of Albanet and related applications. Technical report, University of Vlora. (`http://fjalnet.com/technicalreportalbanet.pdf`).

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Kiril Simov and Petya Osenova. 2010. Constructing of an ontology-based lexicon for Bulgarian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. URL `http://www.lrec-conf.org/proceedings/lrec2010/summaries/848.html`.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, page 781–784. Lisbon.

Liling Tan and Francis Bond. 2013. XLING: Matching query sentence to parallel corpus using topic models for word sense disambiguation. In *International Workshop on Semantic Evaluation (SemEval 2013)*.

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP),*. Suntec, Singapore.

Antonio Toral, Stefania Bracal, Monica Mona-

chini, and Claudia Soria. 2010. Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452. Szeged.

Piek Vossen and Marten Postma. 2014. Open Dutch wordnet. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*. Tartu. (presentation only).

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38. URL `\url{http://doi.acm.org/10.1145/1852102.1852106}`.