Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2984–2990
Marseille, 11–16 May 2020

CA-EHN: Commonsense Analogy from E-HowNet

Peng-Hsuan Li, Tsan-Yu Yang, Wei-Yun Ma
Academia Sinica
jacobvsdanniel@iis.sinica.edu.tw, s0920331239@gmail.com, ma@iis.sinica.edu.tw

Abstract

Embedding commonsense knowledge is crucial for end-to-end models to generalize inference beyond training corpora. However, existing word analogy datasets have tended to be handcrafted, involving permutations of hundreds of words with only dozens of pre-defined relations, mostly morphological relations and named entities. In this work, we model commonsense knowledge down to word-level analogical reasoning by leveraging E-HowNet, an ontology that annotates 88K Chinese words with their structured sense definitions and English translations. We present CA-EHN, the first commonsense word analogy dataset containing 90,505 analogies covering 5,656 words and 763 relations. Experiments show that CA-EHN stands out as a great indicator of how well word representations embed commonsense knowledge. The dataset is publicly available at `https://github.com/ckiplab/CA-EHN`.

Keywords: Corpus, Lexicon, Lexical Database, Ontologies

1. Introduction

Commonsense reasoning is fundamental for natural language agents to generalize inference beyond training corpora. Although the natural language inference (NLI) task (Bowman et al., 2015; Williams et al., 2018) has proved a good pre-training objective for sentence representations (Conneau et al., 2017), commonsense coverage is implicit and limited by the amount of annotated sentence pairs. Furthermore, most models are still end-to-end, relying heavily on word representations to provide background world knowledge.

Therefore, it is desirable to model commonsense knowledge down to word-level analogical reasoning. In this sense, existing analogy benchmarks are lackluster. For Chinese analogy (CA), the simplified Chinese dataset CA8 (Li et al., 2018) and the traditional Chinese dataset CA-Google (Chen and Ma, 2018), translated from English (Mikolov et al., 2013), contain only a few dozen relations, most of which are either morphological, e.g., a shared prefix, or about named entities, e.g., capital-country.

However, commonsense knowledge bases such as WordNet (Miller, 1995) and ConceptNet (Speer and Havasi, 2012) have long annotated relations in our lexicon. Among them, E-HowNet (Chen et al., 2005; Ma and Shih, 2018), extended from HowNet (Dong and Dong, 2003), currently annotates 88K traditional Chinese words with their structured definitions and English translations.

In this paper, we propose an algorithm to extract accurate analogies from E-HowNet with refinements from linguists. We present CA-EHN, the first commonsense analogy dataset containing 90,505 analogies covering 5,656 words and 763 relations. In the experiments, we show that it is useful to embed more commonsense knowledge and that CA-EHN tests this aspect of word embedding.

2. Related Work

In this work, we use word sense definitions from the E-HowNet (Chen et al., 2005; Ma and Shih, 2018) ontology as well as further linguist refinements to construct our commonsense word analogy corpus. Compared to the WordNet (Miller, 1995) gloss, E-HowNet has structured definitions for word senses, each of which can be parsed into a definition graph. These graphs are fundamentally different from that of ConceptNet (Speer and Havasi, 2012). In ConceptNet, there is one huge graph where each node is a concept (words) and each edge is a relation induced by two concepts. For example, there is a **capable-of** edge from **bird** to **fly**. In this work, for each word sense, we create its defining graph, where edges represent modifying attributes. For example, there is a **predication** edge from **animal** to **fly** in the defining graph of **bird**. More detailed and precise descriptions are given in Section 3. and Section 4..

Notable Chinese word analogy datasets include CA8 (Li et al., 2018) and CA-Google (Chen and Ma, 2018). The former is created by Chinese annotators, and the later is translated from English analogies to Chinese. Both of their analogies are essentially the permutation of word pairs that span only a few dozens of relations, mostly regarding named entities and morphology. In contrast, the analogies of CA-EHN are extracted from the E-HowNet lexical ontology and span hundreds of common sense relations. Table 3 shows detailed statistics of these word analogy corpora.

3. E-HowNet

E-HowNet 2.0[1] (Ma and Shih, 2018) consists of two major parts: A lexicon of *words*, *concepts*, and *attributes* (Table 1), and a taxonomy of concepts (Figure 1).

3.1. Lexicon

The E-HowNet lexicon consists of 88K words, 4K concepts, and dozens of attributes. These three types of tokens can be readily distinguished by token names: Words, such as 人 and 雞, are entirely in Chinese. Concepts, such as human|人 and 雞|chicken, contain a vertical bar and an English string in their name. (The order of English and Chinese does not matter in this work.) Attributes, such as telic and theme, are always in English.

In the lexicon, each word is annotated with one or more word senses, and each word sense has a structured *defini-*

[1] 廣義知網2.0版(`http://ehownet.iis.sinica.edu.tw`)

Token	Type	Definitions
telic	attribute	
協	word	#1:{help|幫助} #2:{community|團體}
駿馬|ExcellentSteed	concept	{馬|horse:qualification={HighQuality|優質}}
實驗室	word	#1:{InstitutePlace|場所:telic={or({experiment|實驗:location={~}},{research|研究:location={~}})}}

Table 1: E-HowNet lexicon.

tion. Each definition consists of concepts connected by attributes. Furthermore, every concept also comes with one such structured definition. In essence, words are defined by concepts, and concepts are recursively defined by more abstract concepts. Table 1 shows a part of the lexicon, with gradually more complex definition examples:

- The attribute telic has no definition.
- The word 協 has two senses. The first (help) is trivially defined by help|幫助, and the second (association) by community|團體.
- The concept 駿馬|ExcellentSteed is defined as 馬|horse modified by HighQuality|優質 with the qualification attribute.
- The word 實驗室 has only one sense (laboratory), defined as an InstitutePlace|場所 used as the location for experiment|實驗 or research|研究.

In this work, we use E-HowNet word sense definitions to extract commonsense analogies (Section 4.). Besides, word senses are annotated with their English translations, which could be used to transfer our extracted analogies to English multi-word expressions.

3.2. Taxonomy

The E-HowNet taxonomy organizes concepts into a tree. Figure 1 shows the partially expanded taxonomy. The words beside each node have senses defined trivially by that concept. For example, one definition of 東西 is simply {thing|萬物}.
In the experiments, we infuse E-HowNet taxonomy to distributed word representations and analyze performance changes across word analogy benchmarks (Section 5.4.).

4. Commonsense Analogy

We extract word analogies with rich coverage of words and commonsense relations by comparing word sense definitions (Section 3.1.). The extraction algorithm is further refined with multiple filters and linguist annotations.

4.1. Analogy Extraction

Before refinement, for each sense pair of two words, we try to extract analogies with the following five steps. This process is illustrated in Figure 2.

Figure 1: E-HowNet taxonomy.

Definition Expansion A definition is expanded if it contains only one concept. For example, 駿馬 is defined simply as {駿馬|ExcellentSteed}. Such trivial definitions would only lead to trivial analogies equating synonym pairs. We resolve this problem by replacing the definitions of those words with the definitions of their defining concepts. For example, the definition of 駿馬 is replaced by {馬|horse:qualification={HighQuality|優質}}, i.e., the definition of 駿馬|ExcellentSteed.

Definition Parsing Each definition is parsed into a directed graph. Each node in the graph is either a word, a concept, or a *function*, e.g., or() at the bottom of Table 1. Each edge either links to an attribute modifier, e.g., :telic=, or connects a function node with its argument nodes. Figure 3 shows some more parsed definition graphs.

Graph Comparison The two definition graphs are compared to see if they differ only in one concept node. If they do, the two (word, concept) pairs are analogical to one another. For example, since the graph of 良材 sense#2 (the good timber) and the expanded graph of 駿馬 sense#1 (an excellent steed) differ only in wood|木

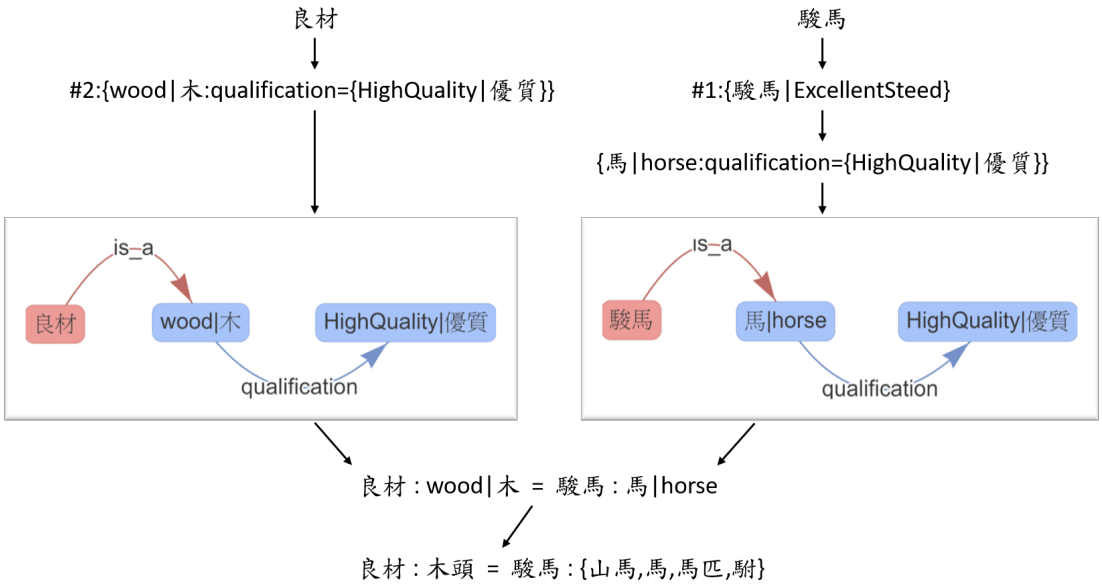
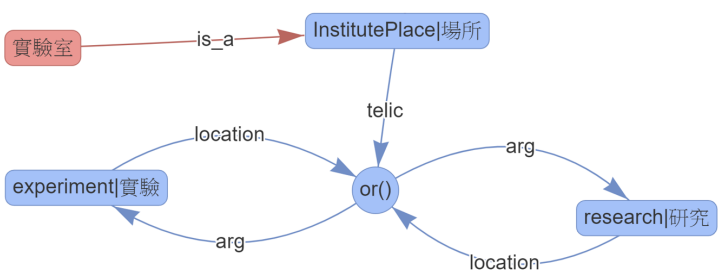
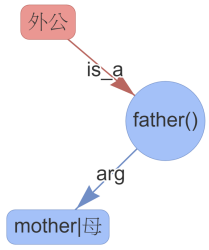


Figure 2: Commonsense analogy extraction.



(a) 實驗室 (laboratory).



(b) 外公 (maternal grandpa).

Figure 3: Sample parsed definition graphs.

and 馬|horse, we extract the following *concept analogy* – 良材:wood|木=駿馬:馬|horse.

Left Expansion The left concept in the concept analogy is expanded into synonym words, i.e., words that have one sense defined trivially by it. For example, there is only one word 木頭 defined as {wood|木}. Thus after expansion, there is still only one analogy: 良材:木頭=駿馬:馬|horse. Most of the time, this step yields multiple analogies per concept analogy.

Right Expansion Finally, the right concept in each analogy is also expanded into synonym words. However, this time we do not use them to form multiple analogies. Instead, the word list is kept as a *synset*. For example, as 山馬, 馬, 馬匹, 駙 all have one sense defined as {馬|horse}, the final analogy becomes 良材:木頭=駿馬:{山馬,馬,馬匹,駙}. The reason why not making multiple analogies in this final step is explained in Section 4.2..

4.2. Embedding Evaluation

Word analogies are typically used for the intrinsic evaluation of word embeddings. For each analogy w_1:w_2=w_3:w_4, the tuple (w_1, w_2, w_3) is called an analogy question and w_4 is the answer. A word embedding must predict the answer as the word of which vector is closest to $v_3 + v_2 - v_1$. In extracting word analogies from E-HowNet, the left expansion step creates plausible analogy questions, but the above embedding evaluation will not work if the right expansion step creates multiple analogies with the same analogy question. This is why the final step keeps the expanded words in a synset. When evaluating embeddings on our benchmark, a predicted word is considered correct as long as it belongs to the synset.

4.3. Accurate Analogy

As the core procedure yields an excessively large benchmark, added to the fact that E-HowNet word sense definitions are sometimes inaccurate, we add refinements to the extraction process to extract a feasible sized benchmark of accurate analogies.

Concrete Concepts At every step of the extraction process, we require every word and concept to be under physical|物質. As shown in Figure 1, this excludes abstract taxa such as event|事件 and abstract|抽象物. Thus it filters out words that are often hard to accurately define. This restriction shrinks the benchmark by 50%.

Common Words At every step of the extraction process, we require words to occur at least five times in ASBC 4.0

滴答 (tick-tock)	時鐘 (clock)	鼕鼕 (rat-tat)	{鼓} (鼓|drum)
聾子 (deaf person)	耳 (ear)	瞎子 (blind person)	{目, 眸子, 眼, 眼眸, 眼睛} (eye|眼)
外公 (maternal grandpa)	母親 (mother)	祖父 (paternal grandpa)	{父, 父親, 爸, 爸爸, 爹, 爹爹, 老子} (father|父)
蝌蚪 (tadpole)	青蛙 (frog)	孑孓 (wriggler)	{斑蚊,蚊, 蚊子, 蚊蟲} (蚊子|mosquito)

Table 2: CA-EHN. (word:word=word:synset)

Benchmark	Language	Type	#analogies	#words	#relations
CA8-Morphological	Simplified	reduplication A (morph.)	2,554	344	3
		reduplication AB (morph.)	2,535	423	3
		semi-prefix (morph.)	2,553	656	21
		semi-suffix (morph.)	2,535	727	41
CA8-Semantic	Simplified	geography (entity)	3,192	305	9
		history (entity)	1,465	177	4
		nature	1,370	452	10
		people (entity)	1,609	259	5
CA-Google	Traditional*	morph., entity, gender	11,126	498	14
CA-EHN	Traditional	commonsense	90,505	5,656	763

Table 3: Analogy benchmarks. *Translated from English.

(Ma et al., 2001), a segmented traditional Chinese corpus containing 10M words from articles between 1981 and 2007. This eliminates uncommon, ancient words or words with synonymous but uncommon, ancient characters. This restriction further shrinks the remaining benchmark by 90%.

Concept Analogy Annotation After introducing the previous two refinements, 36,100 concept analogies are extracted by the graph comparison step. Then, linguists are asked to follow an annotation guideline to label their correctness. 1,000 of these concept analogies are labeled by all four annotators with $\kappa = 0.76$, indicating a high inter-annotator agreement. This step results in 25,010 remaining concept analogies.

Synset Annotation Before concept expansion, every synset needed by the 25,010 concept analogies is checked again to remove words that are not actually synonymous with the defining concept. For example, all words in {花草, 山茶花, 薰衣草, 鳶尾花} are common words and have a sense defined trivially as FlowerGrass|花草. However, the last three (camellia, lavender, iris) are judged by the annotator as not synonyms but hyponyms to the concept. So, the synset will be refined to {花草}. This step also helps eliminate words in a synset that are using their rare senses, as we do not expect embeddings to encode those senses without word sense disambiguation.

5. Analyses

With the proposed extraction algorithm, refinements, and linguists annotations, we collected 90,505 accurate analogies for CA-EHN. Table 2 shows a few samples of the corpus, covering such diverse domains as onomatopoeia, disability, kinship, and zoology. We then compare CA-EHN to existing word analogy datasets shown in Table 3.

5.1. Relation Category

The relations in the datasets can be classified into three categories:

- Morphological relations.
  For example, the shared prefix 周 (week):
  一:周一=二:周二=三:周三=...
  (one : Monday = two : Tuesday = three : Wednesday = ...)
- Named entity relations.
  For example, states to their currencies:
  美國:美元=丹麥:克朗=印度:盧比=...
  (US : dollar = Denmark : krone = India : rupee = ...)
- Commonsense relations.
  For example, the solid-fluid relation:
  冰:水=雪:雨=固體:液體=...
  (ice : water = snow : rain = solid : fluid = ...)

Existing datasets contain mostly morphological (morph.) or named entity (entity) relations. The few exceptions are the nature part of CA8 (Li et al., 2018) and the gender part

Embedding	CA8-Morph.		CA8-Semantic		CA-Google		CA-EHN	
	acc	cov*	acc	cov*	acc	cov*	acc	cov*
GloVe-Small	0.085	6,917	0.296	4,274	0.381	5,367	0.033	90,505
GloVe-Large	0.115	7,379	0.376	5,761	0.437	8,409	0.044	90,505
SGNS-Large	0.178	7,379	0.374	5,761	0.502	8,409	0.051	90,505

Table 4: Embedding benchmarking. *The number of analogy questions covered by an embedding.

+E-HowNet	CA8-Morph.		CA8-Semantic		CA-Google		CA-EHN	
	acc	Δacc	acc	Δacc	acc	Δacc	acc	Δacc
GloVe-Small	0.113	+33%	0.309	+4%	0.391	+3%	0.092	+179%
GloVe-Large	0.137	+19%	0.376	0%	0.418	-4%	0.113	+157%
SGNS-Large	0.180	+1%	0.379	+1%	0.489	-3%	0.127	+149%

Table 5: E-HowNet retrofit benchmarking.

+HIT-Thesaurus	CA8-Morph.		CA8-Semantic		CA-Google		CA-EHN	
	acc	Δacc	acc	Δacc	acc	Δacc	acc	Δacc
GloVe-Small	0.126	+48%	0.340	+15%	0.415	+9%	0.062	+88%
GloVe-Large	0.150	+30%	0.381	+1%	0.437	0%	0.076	+73%
SGNS-Large	0.204	+15%	0.385	+3%	0.502	0%	0.083	+63%

Table 6: HIT-Thesaurus retrofit benchmarking.

of CA-Google (Chen and Ma, 2018). In contrast, CA-EHN fully dedicates as an extensive benchmark for commonsense word reasoning.

5.2. Relation Diversity

For the total number of covered relations, existing datasets span only dozens of pre-defined relations. Their analogies are then created as the permutations of word pair equations. For example, CA8 uses the province-university relation

- 北京:北京大學=南京:南京大學=海南:海南大學=...
  (Beijin : Peking University = Nanjing : Nanjing University = Hainan : Hainan University = ...)

to create more than two hundred analogies.
In contrast, all the 90,505 analogies in CA-EHN are automatically extracted from the E-HowNet ontology and then verified by linguists. Still, we can group word pairs into equivalence classes to see what relations are present in the corpus. For example, we have both

- 樹苗:樹=蝌蚪:青蛙
  (sapling : tree = tadpole : frog)
- 蝌蚪:青蛙=孑孓:蚊
  (tadpole : frog = wriggler : mosquito)

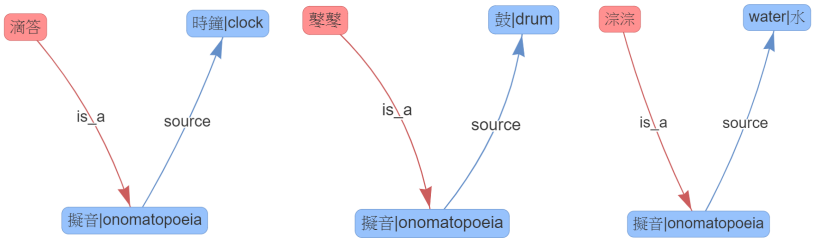
So we can easily know that (樹苗, 樹) and (孑孓, 蚊) belong to the same equivalence class, which seems to express the juvenile-adult relation. By grouping all 90,505 commonsense analogies into equivalence classes, we find that CA-EHN have an unprecedented coverage of 763 relations. Figures 4, 5, 6, 7 show some of the relations.

5.3. Embedding Benchmarking

To evaluate the robustness of using CA-EHN for the classic intrinsic embedding evaluation task (Section 4.2.), we trained and tested different word embeddings across different benchmark datasets.
We trained each word embedding using either GloVe (Pennington et al., 2014) or SGNS (Mikolov et al., 2013) on a small or a large corpus. The small corpus consisted of the traditional Chinese part of Chinese Gigaword (Graff and Chen, 2003) and ASBC 4.0 (Ma et al., 2001). The large corpus additionally included the Chinese part of Wikipedia. When calculating accuracy, only those analogy questions of which all words were in an embedding were considered. So a smaller dictionary was not penalized by lower analogy question coverage.
Table 4 shows the results of different combinations of embeddings and benchmarks. It can be seen that CA-EHN is a robust benchmark for the analogy task. On all existing benchmarks and CA-EHN, it is consistent that GloVe-Small is the worst-performing and SGNS-Large is the best. Furthermore, the new dedicated commonsense analogy corpus appears substantially more challenging than existing analogies for distributed word representations.

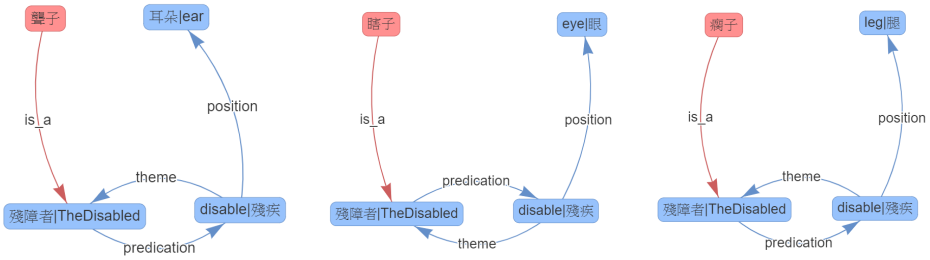
5.4. Commonsense Benchmarking

Two central hypotheses of this work are that it is useful to embed more commonsense knowledge and that CA-EHN tests this aspect of word embedding. To verify these hypotheses, we infused some structure knowledge of commonsense ontology to word embeddings and observed their performance change across benchmarks.



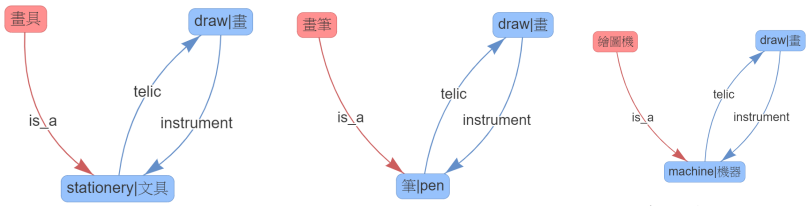
(a) 滴答 (tick-tock). (b) 鼕鼕 (rat-tat). (c) 淙淙 (gurgle).

Figure 4: Some definition graphs that leads to the sound-origin relation.



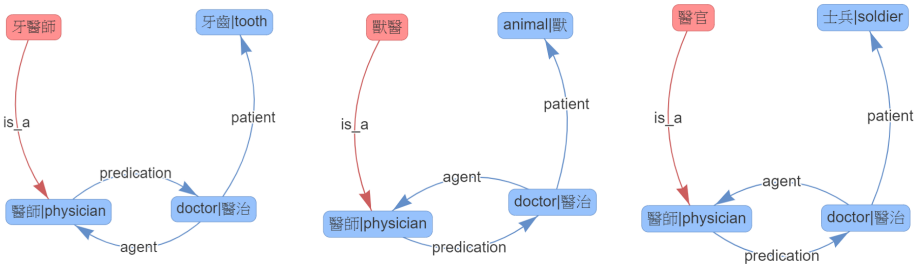
(a) 聾子 (deaf person). (b) 瞎子 (blind person). (c) 瘸子 (lame person).

Figure 5: Some definition graphs that leads to the organ-disabled relation.



(a) 畫具 (painting tools). (b) 畫筆 (painting brush). (c) 繪圖機 (plotter).

Figure 6: Some definition graphs that leads to the painter-instrument relation.



(a) 牙醫 (dentist). (b) 獸醫 (veterinarian). (c) 醫官 (medical officer).

Figure 7: Some definition graphs that leads to the doctor-patient relation.

We infused distributed word representations with the hypo-hyper and same-taxon knowledge in the E-HowNet taxonomy (Section 3.2.) and the HIT-Thesaurus[2] through retrofitting (Faruqui et al., 2015). For example, in Figure 1, the word vector of 物體 was optimized to be close to both its distributed representation and the word vectors of 物質 (same-taxon) and 東西 (hypo-hyper).

Table 5, 6 shows the results of different combinations of retrofitted embeddings and benchmarks. Firstly, retrofitted embeddings achieve better performance on most existing datasets, suggesting the benefits of embedding more commonsense knowledge. Secondly, on CA-EHN, each retrofitted embedding significantly outperforms its pure distributed counterpart in Table 4. Performance increases by up to 179% and 88% by infusing E-HowNet taxonomy and HIT-Thesaurus respectively. This shows that CA-EHN is a great indicator of how well word representations embed commonsense knowledge.

[2] 同義詞詞林擴展版(https://github.com/taozhijiang/chinese_correct_wsd)

6. Conclusion

We have presented CA-EHN, a large and dedicated commonsense word analogy dataset, by leveraging word sense definitions in E-HowNet. After linguist checking, we have 90,505 Chinese analogies covering 5,656 words and 763 commonsense relations. The experiments showed that CA-EHN could become an important benchmark for testing how well future embedding methods capture commonsense knowledge, which is crucial for models to generalize inference beyond their training corpora. With translations provided by E-HowNet, Chinese words in CA-EHN can be transferred to English multi-word expressions.

7. Acknowledgements

We are grateful for the insightful comments from anonymous reviewers. This work is supported by the Ministry of Science and Technology of Taiwan under grant numbers 109-2634-F-001-010, 109-2634-F-001-008.

8. Bibliographical References

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

9. Language Resource References

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Chen, C.-Y. and Ma, W.-Y. (2018). Word embedding evaluation datasets and wikipedia title embedding for chinese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Chen, K.-J., Huang, S.-L., Shih, Y.-Y., and Chen, Y.-J. (2005). Extended-HowNet: A representational framework for concepts. In *Proceedings of OntoLex 2005 - Ontologies and Lexical Resources*.

Dong, Z. and Dong, Q. (2003). Hownet - a hybrid language and knowledge resource. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*.

Graff, D. and Chen, K. (2003). Chinese gigaword ldc2003t09. *Linguistic Data Consortium*.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., and Du, X. (2018). Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Ma, W.-Y. and Shih, Y.-Y. (2018). Extended hownet 2.0 – an entity-relation common-sense representation model. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Ma, W.-Y., Hsieh, Y.-M., Yang, C.-H., and Chen, K.-J. (2001). Design of management system for chinese corpus construction. In *Proceedings of Research on Computational Linguistics Conference XIV*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.