Semantic Interpretation of Superlative Expressions via Structured Knowledge Bases

Sheng Zhang[1], **Yansong Feng**[1*], **Songfang Huang**[2], **Kun Xu**[1], **Zhe Han**[1] and **Dongyan Zhao**[1]
[1]ICST, Peking University, Beijing, China
[2]IBM China Research Lab, Beijing, China
{evancheung, fengyansong, xukun, zhehan1992, zhaodongyan}@pku.edu.cn
huangsf@cn.ibm.com

Abstract

This paper addresses a novel task of semantically analyzing the comparative constructions inherent in attributive superlative expressions against structured knowledge bases (KBs). The task can be defined in two-fold: first, selecting the comparison dimension against a KB, on which the involved items are compared; and second, determining the ranking order, in which the items are ranked (ascending or descending). We exploit Wikipedia and Freebase to collect training data in an unsupervised manner, where a neural network model is then learnt to select, from Freebase predicates, the most appropriate comparison dimension for a given superlative expression, and further determine its ranking order heuristically. Experimental results show that it is possible to learn from coarsely obtained training data to semantically characterize the comparative constructions involved in attributive superlative expressions.

1 Introduction

Superlatives are fairly common in natural languages and play an essential role in daily communications, when in conveying comparisons among a set of items or degrees of certain properties. Properly analyzing superlative expressions holds the promise for many applications such as question answering (QA), text entailment, sentiment analysis and so on. In literature, analysis of superlatives has drawn more interests from both formal linguistics and semantics(Szabolcsi, 1986; Gawron, 1995; Heim, 1999; Farkas and Kiss, 2000), but relatively less attention from the computational linguistics and NLP communities(Bos and Nissim, 2006; Jindal and Liu, 2006; Scheible, 2007; Scheible, 2009).

Earlier computational treatments to superlatives focus on the categorizations of superlatives(Bos and Nissim, 2006; Scheible, 2009), where the most common but important type is being part of a noun phrase or describing certain properties or attributes of the subjects, named as *attributive superlatives* or *ISA-superlatives*, accounting for around 90% of appearances in newswire. A typical example is *Nile is the longest river in the world*.

In most cases, the gist behind such superlative expressions is the comparative constructions that the utterance intends to convey to readers, e.g., in the *Nile* example, *Nile* is *longer* than any other rivers in the world. Semantically understanding such attributive superlative expressions boils down to interpreting the comparative construction involved in the utterance in the following four aspects:

1. **Target**: one or more items that work as the protagonist of the utterance, and are being compared within the comparative construction, e.g., *Nile*;
2. **Comparison set**: the set of items that are being compared against in the utterance(Bos and Nissim, 2006), e.g., *all rivers in the world*;
3. **Comparison dimension**: the attribute or property that the items are compared upon, e.g., *the length of a river*;
4. **Ranking order**: the order in which the items in the comparison set are sorted according to the dimension, in an ascending or descending order, e.g., we should rank all rivers regarding their lengths in a *descending* order to get the *longest* at the top.

So far, there have been only a few computational treatments for superlatives, addressing the

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 225–230,
Beijing, China, July 26-31, 2015.

importance of categorizing superlatives, identifying the target and comparison set(Bos and Nissim, 2006; Jindal and Liu, 2006; Scheible, 2009), while putting less attention on other aspects.

In fact, grounding the comparison dimension into a canonical predicate of a KB can help provide more accurate interpretations for the involved comparative constructions. In question answering over structured KBs, accurate treatments for superlatives will not only help build more precise of structured queries, but also support shallow functional reasoning, e.g., formally analyzing *the fifth longest river in the world*, will explore the most of the structured nature of KBs, and is advantageous to traditional IR based methods.

However, selecting an appropriate comparison dimension against a structured KB is not a trivial task. Usually, the numbers of adjective superlatives and gradable KB predicates are large, so that it is impossible to craft mapping rules to cover every pair of adjective superlative and predicate. Consequently, preparing wide-coverage annotated data to help automate this procedure is also labor-intensive and time consuming. Moreover, some adjectives are widely used, but often vague to decide a dimension by themselves(Bos and Nissim, 2006). One may need to draw support from their context and even common sense knowledge.

In this paper, we propose a novel task, semantically interpreting the comparative constructions inherent in attributive superlative expressions again structured KBs, e.g., Freebase(Google, 2013), specifically, focusing on selecting appropriate comparison dimensions and corresponding ranking orders. To this end, we collect training data from roughly aligned Wikipedia resources and knowledge facts in Freebase, from which we build a neural network model to reveal the underlying correspondence between the comparative construction, as well as its context, and a Freebase predicate. Our method leverages the potentials of structured KBs and large amount of text resources in Wikipedia without relying on human annotated data.

We evaluate our interpretation of superlatives in two tasks, and experimental results show that it is possible to learn the comparison dimensions in form of canonical knowledge bases predicates from roughly collected training data, which is noisy in nature but provides the essentials to semantically characterize superlative expressions.

2 The Task

Given a sentence with an attributive superlative expression, our task is to find on which dimension the comparison happens against a KB and how the comparison results are arranged, i.e., (1) **Dimension Selection:** decide the *dimension* on which the involved items are compared, and ground the selected dimension into Freebase predicates. (2) **Ranking Order Determination:** given the comparison set and the selected dimension, determine the *order* in which the involved items are ranked within the comparisons, in an **ascending** or **descending** order? For superlatives coupled with ordinals, we also need to assign the standing in the rankings.

In the *Nile* example, we expect to interpret *the longest river* into a vector <*fb:geography.river.length*, *descending*, **1**>, where all *rivers in the world* are compared upon Freebase predicate *fb:geography.river.length*, sorted in a *descending* order and the referred target ranks the ***first***.

3 The WikiDiF Dataset

Previous superlative datasets are built to facilitate superlative extraction, classification, and comparison set identification(Bos and Nissim, 2006; Scheible, 2012). There are currently no available datasets that can be used directly for our task, especially no annotations against structured KBs.

We therefore present a distantly supervised method to collect annotated training data from rich text resources of Wikipedia and the help of Freebase, without much human involvement. The **key assumption** behind our method is that if a superlative expression frequently appears in a context that may describe a KB predicate, then this predicate probably plays an important role in the comparative construction triggered by this superlative. Inspired by recent advances in relation extraction(Mintz et al., 2009), given a Freebase predicate, we are able to collect many sentences from Wikipedia pages, which more or less describe this predicate, without extra human annotation. These sentences in turn can be used to collect the co-occurrences between a superlative expression and this predicate.

In more detail, we first find all Freebase predicates that may involve in comparative constructions, i.e., all *gradable* predicates, e.g., *fb:geography.river.length*, on which different rivers can be compared with each other. In practice, we simply treat all Freebase predicates that take objects of type $\in$ {*/type/int*, */type/float*, */type/datetime*} as *gradable* predicates. In total, we collect 8,968,383 <*subj, rel, obj*> triples covering all 1,795 *gradable* predicates from Freebase dump.

Next, we extract, from Wikipedia pages, all sentences containing superlative expressions, as well as their ± 3 context sentences[1]. To detect superlatives, we rely on part-of-speech tags (*JJS, RBS*) which can achieve a high recall in practice according to (Jindal and Liu, 2006). By doing so, we collect 7,734,006 sentences with superlative expressions from Wikipedia.

Finally, for each collected triple <*subj, rel, obj*> , we match *subj* and *obj* into our sentence collection, including those contextual sentences. This gives us 20,609 sentences with superlative expressions that potentially describe our collected knowledge triples with *gradable* predicates. For example, the following sentences from the page of *Nile* in Wikipedia may describes a Freebase fact <*Nile, fb:geography.river.length, 6,853*>:

> "*The Nile is a major north-flowing river in northeastern Africa, generally regarded as the longest river in the world. It is 6,853 km (4,258 miles) long.* "

where we can see that *longest* has implied a comparative construction among all *rivers in the world* and *fb:geography.river.length* is the involved hidden comparison dimension.

Our resulting dataset, WikiDiF, contains **20,609** sentences paired with Freebase predicates, covering **2,335** superlative words and **340** Freebase predicates[2]. In WikiDiF, there are on average 8.8 sentences per superlative word targeting for about 2 predicates, and for commonly seen superlatives, e.g., largest or biggest, there are on average 70 sentences per superlative word targeting for 30 predicates. Compared to other human annotated datasets, WikiDiF is admittedly noisy in nature, but exploits the underlying connections between knowledge facts and their possibly corresponding textual descriptions, where the pseudo co-occurrences of superlative expressions and Freebase predicates will work as a proxy for us to formally analyze the involved comparative constructions.

4 Comparison Dimension Selection

Our WikiDiF dataset contains utterances with roughly annotated *superlative-predicate* pairs, which helps us to model the dimension selection task as a classification problem. Given a superlative word S and its context C, our goal is to find a *gradable* Freebase predicate R that maximizes the conditional probability $P(R \mid C, S)$:

$$R^* = \arg\max_{R \in cand_S} P(R \mid C, S)$$

where we can limit our search space to a candidate set $cand_S$ according to the domain and type constraints regarding the comparison set.

Currently, our WikiDiF covers limited predicates and training instances for each superlative S, traditional classification models may suffer from the coverage and data sparsity issues. Here, we adopt a classic one-layer neural network (NN) model with the help of word embeddings to predict how likely a predicate R can work as the comparison dimension given S and its context vector.

We start by constructing vector representations of words and store them in a table L. We use the publicly available word embeddings trained by SENNA (Collobert et al., 2011), with 50 dimensions throughout our experiments.

We construct the context vector for each instance by concatenating the vectors of context words within a window of $\pm k$. If there are not enough words within the window, special filling vectors will be used.

$$V = (w_{j_S - k}, ..., w_{j_S - 1}, w_{j_S + 1}, ..., w_{j_S + k})$$

where j_S is the index of superlative S in the sentence, and w is the vector of context word or special filling parameter.

The output layer of the NN model is a standard *softmax* function, which takes a *sigmoid* nonlinearity of the context vectors as input. Therefore, the probability of ith predicate in Freebase is chosen as the comparison dimension given superlative S and its context C can be written as:

$$P(R_i \mid C, S) = \frac{e^{z_i}}{\sum_n^{q=1} e^{z_q}}$$

[1] In Wikipedia, sentences with superlative expressions may not always contain the knowledge facts that support the superlative constructions, which often appear in their neighbouring sentences. For example, *highest*, and its supporting fact, (*Everest*, *8,848 metres*), are not in the same sentence, but indeed very near: *Mount Everest, also known ..., is Earth's* ***highest*** *mountain. It is located in ... of the Himalayas. Its peak is* ***8,848 metres*** *(29,029 ft) above sea level.*

[2] We also filter out predicates whose objects are very common in the documents, e.g., 1 or 2, which is difficult to collect training data.

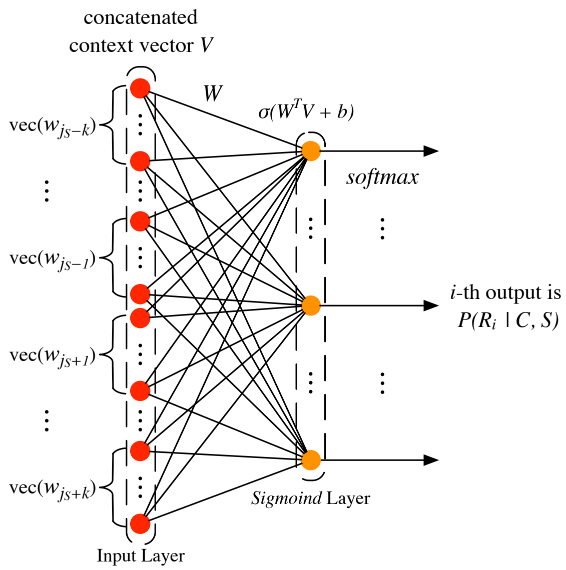

Figure 1: Neural network for dimension selection.

where z_i is estimated using a *sigmoid* function:

$$z_i = \sigma_i(W^T_{m\times n}V + b)$$

where n is the number of candidate predicates for superlative S, m is the length of concatenated vector V, $W_{m\times n}$ is the parameter matrix, b is the bias vector, and σ_i is the *sigmoid* function that applies to the i-th element of argument vector.

Training this standard one-layer neural network model can be straightforward, and the parameters W, b and filling vectors can be updated using stochastic gradient descent (SGD).

5 Ranking Order Determination

To exploit the most from a structured KB, we use an effective heuristic method to decide the ranking order when a superlative expression S triggers a comparative construction regarding a KB predicate R. For each pair of (S, R), we first find from R's supporting sentences the ones that contain superlative S, and further trace the $<subj^*, R, obj^*>$ tuples from the KB which these sentences are assumed to describe. We next look up into the KB, and find other tuples $<subj^o, R, obj^o>$ with the same predicate R. If $subj^o$ and $subj^*$ are of the same type, and most $obj^o < obj^*$, we will say the ranking order for expression S is **descending** when implying predicate R, otherwise, **ascending** order.

In the *Nile* example, if we find in our KB that nearly all other entities of type *river* have smaller values than *6,853* in *fb:geography.river.length* , we can conclude that the ranking order for *longest* regarding *fb:geography.river.length* is *descending* , i.e., we should rank all rivers *descendingly* to get the *longest* one at the top.

Ordinals are processed as a post-processing step to interpret ordinals into a numerical values.

6 Experiments

The main purposes of this work is to answer the following two questions: (1) can we learn from noisy training data without much human involvement to semantically interpret attributive superlative expressions via structured KBs? (2) can our semantic analysis help better understand utterances with superlative expressions?

6.1 Interpreting Superlative Comparisons

We first evaluate our models in the vanilla setup defined in Section 2[3], in terms of accuracy of dimension selection (Acc_d), precision of predicates covered by WikiDiF (P_d), and precision of ranking order determination (P_o).

Datasets: We manually annotate superlative expressions from QALD-4 evaluation dataset(Unger et al., 2014) and TREC QA (2002, 2003) datasets(NIST, 2003), and guarantee that all the labeled superlative instances can be grounded to gradable Freebase predicates. The resulting question dataset contains 135 questions covering 44 Freebase predicates. Additionally, we manually annotate 77 declarative sentences covering 24 Freebase predicates from WSJ and Wikipedia as the declarative dataset.

We build a Naïve Bayes model using co-occurrences as a baseline to predict proper *dimensions*. We further implement a simple baseline to decide the ranking order, by measuring the relatedness between a superlative word and two sets of seed words using word embeddings. The two seed sets, {*most, more, much, many*} and {*least, less, few, little*}, potentially indicate two ranking orders, respectively.

We can see in Table 1 that our method can better capture the underlying connections between superlative expressions and KB predicates. Comparing with the baseline, our model benefits from the NN architecture and distributional word representations and avoids data sparsity to some ex-

[3]We assume that the domain and type constraints regarding the comparison set are known

Model	Questions			Declaratives		
	Acc_d	P_d	P_o	Acc_d	P_d	P_o
Baselines	40.7	81.7	95.5	25.9	78.5	97.4
Ours	48.9	92.9	99.2	33.8	92.8	100

Table 1: Performances of superlative interpretations on two datasets

tent. Our model achieves over 90% of precision on the predicates covered by our training data, showing that it is possible to learn from noisy training data to characterize comparison dimensions against a KB. However, the relatively lower accuracy is mainly due to that some predicates in the testing data are not covered by our WikiDiF, which only covers 19% of gradable Freebase predicates. Regarding the ranking order determination, our simple heuristic method makes the most of Freebase triples, and outperforms the relatedness based baseline, which does not take Freebase into account.

By looking at the different performances on question and declarative sentences, we can see that our method performs better on relatively simpler and shorter questions, while a slightly worse on longer declarative sentences. This is not surprising, since questions are often simple in structure and ask for a straightforward property about the target, while declarative sentences are usually complicated in syntax, which a ±5 context words window may not be able to capture. Another reason is that for newswire, there are many predicates that are similar in definitions but with tiny differences, which often confuse our methods, e.g., *fb:business.business_operation.assets* and *fb:business.business_operation.current_assets*.

6.2 Question Answering over Freebase

We also investigate how our semantic analysis for superlatives can help improve question answering on two benchmark datasets, Free917(Cai and Yates, 2013) and WebQuestions(Berant et al., 2013), which contain 35 questions with attributive superlative expressions in total. We inject our formal analysis as superlative-triggered aggregation operations into an existing system, Xu14(Xu et al., 2014). Note that we leave the *comparison_set* to be decided by Xu14's parser.

The 35 superlative-triggered complex questions can not be correctly answered by most state-of-the-art systems(Berant et al., 2013; Yao and Van Durme, 2014; Xu et al., 2014), since they can not properly analyze the superlative-triggered functions. When integrated with our analysis, Xu14 is able to correctly answer **14 out of 35** such questions (**40%**), significantly outperforming other systems. The remaining 21 questions are mainly idiomatic usage, e.g., *the Best Actor Award*, or with predicates not covered by WikiDiF.

The result shows that our analysis can help QA systems better handle superlative-triggered aggregation functions, which previous works fail to do. This also gives a good reason to introduce the analysis for comparison constructions into the QA community, which will leverage the potentials of structured KBs to better deal with complex questions.

7 Conclusion

In this paper, we present a novel attempt to semantically analyze the comparisons involved in attributive superlative expressions by investigating on which dimension the comparative construction works and how the comparison results are arranged. We leverage Freebase and their roughly aligned textural descriptions from Wikipedia, and learn from such training data to characterize a comparative construction in two aspects, the dimension of the comparison and its ranking order.

Currently, our analysis suffers from the limited coverage of our WikiDiF. In the future, it would be interesting to improve our method to cover more KB predicates, and extend our NN model with more advanced structures to further improve the performances and also simultaneously characterize the target and comparison set involved.

Acknowledgments

We would like to thank Heng Ji, Benjamin Van Durme, Liwei Chen and Bingfeng Luo for their helpful discussions and three anonymous reviewers for their insightful comments that greatly improved the work. This work was supported by National High Technology R&D Program of China (Grant No. 2015AA015403, 2014AA015102), Natural Science Foundation of China (Grant No. 61202233, 61272344, 61370055) and the joint project with IBM Research. Any correspondence please refer to Yansong Feng.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544.

Johan Bos and Malvina Nissim. 2006. An empirical approach to the interpretation of superlatives. In *EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 9–17.

Qingqing Cai and Alexander Yates. 2013. Semantic Parsing Freebase: Towards Open-domain Semantic Parsing. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Donka F Farkas and Katalin É Kiss. 2000. On the comparative and absolute readings of superlatives. *Natural Language & Linguistic Theory*, 18(3):417–455.

Jean Mark Gawron. 1995. Comparatives, superlatives, and resolution. *Linguistics and Philosophy*, 18(4):333–380.

Google. 2013. Freebase data dumps. `https://developers.google.com/freebase/data`.

Irene Heim. 1999. Notes on superlatives.

Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1331–1336.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011.

NIST. 2003. Text retrieval conference question answering track.

Silke Scheible. 2007. Towards a computational treatment of superlatives. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, ACL '07, pages 67–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Silke Scheible. 2009. *A Computational Treatment of Superlatives*. Ph.D. thesis, University of Edinburgh.

Silke Scheible. 2012. Textwiki: a superlative resource. *Language Resources and Evaluation*, 46(4):635–666.

Anna Szabolcsi. 1986. Comparative superlatives. *MIT Working papers in Linguistics*, 8:245–265.

Christina Unger, Corina Forascu, Vanessa Lopez, Axel Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. 2014. Question answering over linked data (qald-4). In *Proceedings of the CLEF 2014*, pages 1172–1180.

Kun Xu, Sheng Zhang, Yansong Feng, and Dongyan Zhao. 2014. Answering natural language questions via phrasal semantic parsing. In *Proceedings of the 2014 Conference on Natural Language Processing and Chinese Computing (NLPCC)*, pages 333–344.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of ACL*.