Implementing Natural Language Inference for Comparatives

Department of Computer Science Izumi Haruta (Supervisor : Daisuke Bekki)

1 Introduction

Comparative constructions pose a challenge in Natural Language Inference (NLI), which is the task of determining whether a text entails a hypothesis. In formal semantics, there is a rich body of work on comparatives and gradable expressions using the notion of degree. However, a logical inference system for comparatives has not been sufficiently developed for use in the NLI task. In this paper, we present a compositional semantics that maps various comparative constructions in English to logical forms (LFs) via Combinatory Categorial Grammar (CCG) parsers and combine it with an inference system based on automated theorem proving.[1]

2 System architecture

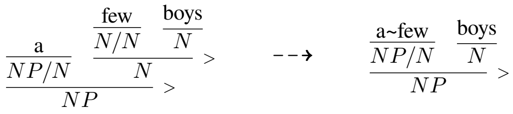
Figure 1 shows the pipeline of the proposed system. First, the premises and hypothesis are mapped to LFs based on A-not-A analysis via CCG parsing and tree transformation. Next, a theorem prover judges *yes*, *no*, or *unknown* with the axioms for comparatives and lexical knowledge.

Degree semantics: A-not-A analysis To analyze gradable adjectives, we use the two-place predicate of entities and degrees as developed in degree-based semantics [1]. For instance, the sentence *Ann is 5 feet tall* is analyzed as $\mathsf{tall}(\mathsf{ann}, 5\ \mathsf{feet})$, where $\mathsf{tall}(x, \delta)$ is read as "x is (at least) as tall as degree δ". Comparative expressions are analyzed in terms of first-order logic, using the so-called A-not-A analysis [2]; for example, *Chris is taller than Alex* is analyzed as $\exists\delta(\mathsf{tall}(\mathsf{chris}, \delta) \wedge \neg\mathsf{tall}(\mathsf{alex}, \delta))$, which asserts that there is a degree δ of tallness that Chris satisfies but Alex does not. We present semantic LFs for some example sentences using A-not-A analysis, as shown in Table 1. This analysis can be naturally extended to not comparative forms of adjectives such as (1), but also generalized quantifiers, adverbial phrases, and the comparative forms of adverbs such as (2-4).

Compositional semantics in CCG In CCG, the mapping from syntax to semantics is defined by assigning syntactic categories to words [3]; the LF of a sentence is then compositionally derived using λ-calculus. However, there is a gap between the syntactic structures assumed in formal semantics and the output derivation trees of existing CCG parsers. For this reason, we modify the derivation trees provided by CCG parsers in post-processing. For instance, we compound expressions for comparatives and quantifiers are combined as one word, such as *a few, a lot of*, and *at most.*

Table 1: Logical forms of gradable constructions based on A-not-A analysis

Sentence	Logical form
(1) Ken is *2 inches taller than* Harry.	$\forall\delta(\mathsf{tall}(\mathsf{harry}, \delta - 2'') \rightarrow \mathsf{tall}(\mathsf{ken}, \delta))$
(2) *Most* apples are red.	$\exists\delta(\exists x(\mathsf{apple}(x) \wedge \mathsf{red}(x) \wedge \mathsf{many}(x, \delta))$ $\wedge\neg\exists x(\mathsf{apple}(x) \wedge \neg\mathsf{red}(x) \wedge \mathsf{many}(x, \delta)))$
(3) *Few* children ran.	$\neg\exists x\exists\delta(\mathsf{child}(x) \wedge \mathsf{many}(x, \delta) \wedge (\delta > \theta_{\mathsf{many}}(\mathsf{child}))$ $\wedge\exists e(\mathsf{run}(e) \wedge (\mathsf{subj}(e) = x)))$
(4) Bob drove *as carefully as* John.	$\exists e_1\exists e_2(\mathsf{drive}(e_1) \wedge (\mathsf{subj}(e_1) = \mathsf{bob}) \wedge \mathsf{drive}(e_2)$ $\wedge(\mathsf{subj}(e_2) = \mathsf{john}) \wedge \forall\delta(\mathsf{careful}(e_2, \delta) \rightarrow \mathsf{careful}(e_1, \delta)))$



Insertion of lexical knowledge To test the compatibility of logical inferences and inferences involving lexical knowledge, we implement an abduction mechanism to search for useful axioms drawn from knowledge bases before the process of theorem proving. The strategy is similar to the one used in previous studies [4] in which the system searches for lexical relations from WordNet [5] and VerbOcean [6].

3 Experiments

For evaluation, we use five datasets: FraCaS [7]; MED [8]; SICK [9]; HANS [10]; CAD. CAD was created for this study and includes problems related to pragmatic inference.

Table 2 gives the results of the evaluation. *Maj* is the accuracy of the majority baseline and *Ours* the accuracy of our system. For FraCaS, MED, and CAD, +*rule* shows the accuracy achieved by the addition of hand-coded rules, which correct the errors in POS tagging and lemmatization. For CAD, we also experimented with an implementation for pragmatic inference. The accuracy is shown in +*imp*.

We compared our system with four logic-based systems (`MN` [11], `LP` [12, 13], `MG` [4], and `GKR4` [14]), three DL-based systems (`RB`, `BERT` [8], `BERT+` [8], and `BF` [15]), and two hybrid systems (`HNB` [14] and `HNX` [14]). `RB` is a system that used a state-of-the-art model, RoBERTa (`RB`) [16], trained on MultiNLI [17].

The results show that our system achieved high accuracy on the logical inferences with adjectives, comparatives, quantifiers, numerals, and adverbs. For

[1] GitHub repository with code and data: `https://github.com/izumi-h/ccgcomp`

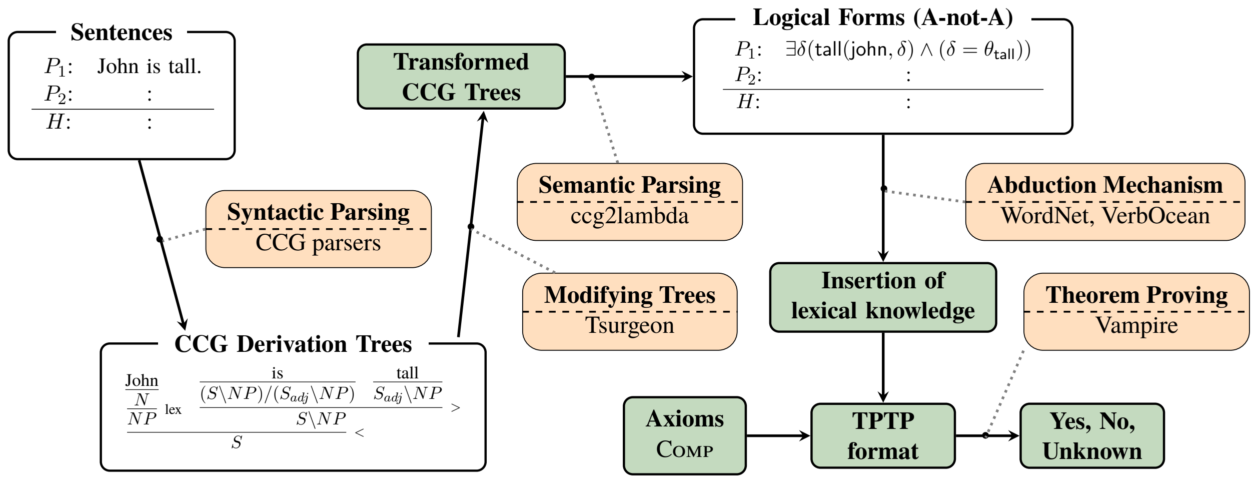


Figure 1: Overview of the proposed system

Table 2: Accuracy on FraCaS, MED, SICK, HANS, and CAD datasets

FraCaS				
Section	*GQ*	*Adj*	*Com*	*Att*
#All	73	22	31	13
Maj	.49	.41	.61	.62
RB	.73	.45	.52	.69
MN	.77	.68	.48	.77
LP	.93	.73	–	**.92**
Ours	.97	.82	.90	.92
+rule	**.99**	**.95**	**.90**	**.92**

MED		
Label	*gq*	*gqlex*
#All	498	691
Maj	.58	.63
BERT	.56	.58
BERT+	.54	.68
RB	.57	.55
Ours	**.97**	.91
+rule	**.97**	**.92**

SICK	
#All	4927
Maj	.57
RB	.56
LP	.81
MG	**.83**
Ours	.82

HANS		
Gold	*yes*	*unknown*
#All	15000	15000
Maj	.50	.50
BF	.87	.61
RB	**1.0**	.56
GKR4	.84	.59
HNB	.84	.54
HNX	.83	.25
Ours	.98	**.83**

CAD	
#All	257
Maj	.43
RB	.58
Ours	.81
+rule	.82
+rule +imp	**.92**

HANS, [10] reported that DL-based systems tend to erroneously output *yes* for cases in which the hypothesis was a constituent or a sub-string of the premise, such as disjunctive sentences. To see how a system performs on these cases, we present the accuracy for each gold answer label (*yes* and *unknown*). While the accuracy when the gold label was *yes* was close to 100% in both our system and the DL-based system (RB), the accuracy of our system outperformed RB when the label is *unknown*.

4 Conclusion

In this study, we presented an end-to-end logic-based inference system for handling complex inferences with comparatives, quantifiers, numerals, and adverbs. The entire system is transparently composed of several modules and can solve complex inferences for the right reason. This study contributes to the study of computational modeling and the evaluation of formal semantic theories, as well as to the creation of challenging NLI problems that DL-based models need to address.

References

[1] M. J. Cresswell. The semantics of degree. In Barbara Partee, editor, *Montague Grammar*, pp. 261–292. Academic Press, 1976.

[2] E. Klein. The interpretation of adjectival comparatives. *Journal of Linguistics*, Vol. 18, No. 1, pp. 113–136, 1982.

[3] M. Steedman. *The Syntactic Process*. MIT press Cambridge, MA, 2000.

[4] P. Martínez-Gómez, K. Mineshima, Y. Miyao, and D. Bekki. On-demand injection of lexical knowledge for recognising textual entailment. In *In Proc. of EACL*, pp. 710–720, 2017.

[5] G. A. Miller. WordNet: A lexical database for english. *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[6] T. Chklovski and P. Pantel. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proc. of EMNLP*, pp. 33–40, 2004.

[7] R. Cooper, R. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspers, H. Kamp, M. Pinkal, M. Poesio, S. Pulman, et al. FraCaS–a framework for computational semantics, 1996.

[8] H. Yanaka, K. Mineshima, D. Bekki, K. Inui, S. Sekine, L. Abzianidze, and J. Bos. Can neural networks understand monotonicity reasoning? In *In Proc. of ACL Workshop*, pp. 31–40, 2019.

[9] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *In Proc. of LREC*, pp. 216–223, 2014.

[10] T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *In Proc. of ACL*, pp. 3428–3448, 2019.

[11] K. Mineshima, P. Martínez-Gómez, Y. Miyao, and D. Bekki. Higher-order logical inference with compositional semantics. In *In Proc. of EMNLP*, pp. 2055–2061, 2015.

[12] L. Abzianidze. A tableau prover for natural logic and language. In *Proc. of EMNLP*, pp. 2492–2502, 2015.

[13] L. Abzianidze. Natural solution to FraCaS entailment problems. In *In Proc. of *SEM*, pp. 64–74, 2016.

[14] A.-L. Kalouli, R. Crouch, and V. de Paiva. Hy-NLI: a hybrid system for natural language inference. In *In Proc. of COLING*, pp. 5235–5249, 2020.

[15] Y. Yaghoobzadeh, R. Tachet, T.J. Hazen, and A. Sordoni. Robust natural language inference models with example forgetting. *arXiv preprint arXiv:1911.03861*, 2019.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized bert pretraining approach, 2019.

[17] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *In Proc. of NAACL-HLT*, pp. 1112–1122, 2018.