4Q1-4in2

Grounded Noun/Verb-Phrases to Images for RTE

韓　丹　　マルティネス・ゴメス　パスクアル

人工知能研究センター, 産業技術総合研究所
Artificial Intelligence Research Center, AIST

Semantic interpretation plays an important role in many natural language processing tasks. One of the challenges is to represent the semantics of the components (i.e. words and phrases). Traditional methods offer limited coverage or fail to recognize the semantic relations. We propose an algorithm to ground noun phrases and verb phrases in images, and deduce the relations between phrases by examining the similarity of two sets of images according to a simple similarity measure and a threshold. We evaluate our method in the task of recognizing textual entailment.

1. Introduction and Related Work

Recognizing textual entailment (RTE) is a task where, given a text T (set of sentences) and a hypothesis H, the objective is to recognize whether T entails H. For example: (T) Some men walk in the tall and green grass. (H) Some people walk in the field. Although humans can easily solve these problems, machines face great difficulties.

The RTE problem has been approached from different perspectives, ranging from purely statistical systems [Lai and Hockenmaier, 2014], to purely logical [Mineshima et al., 2015] and hybrid systems [Beltagy et al., 2013]. We evaluate our idea on top of a logic system, since these systems generally offer a high precision and interpretability, which is useful to our purposes.

In this approach, there are two main challenges. One is to model the logical semantic composition of sentences, guided by the syntax and logical words (e.g. “most”, “not”, “some”, “every”). Another one is to introduce lexical knowledge that describes the relationship between words or phrases (e.g. “men” → “people”, “tall and green grass” → “field”).

Whereas the relationship “men” → “people” can be found in high precision ontological resources, phrasal relations such as “tall and green grass” → “field” are not available in databases despite their size. Moreover, although distributional similarity models have an infinite domain (given a compositional function on words), they fail to identify phrasal entailments (e.g. *guitar* has a high similarity to *piano*, but they do not entail each other). Therefore, we propose to add multi-modal features to classifiers to increase their performance.

In terms of multi-modal semantics, one of the most related works to ours is that of [Kiela and Bottou, 2014]. They constructed multi-modal representations of words by concatenating visual features with pre-trained word vectors. However, our work aims to integrate the idea of phrase image groundings in the downstream application of RTE, which requires additional considerations.

Our contribution is a framework to ground phrases to their visual representations, and judge phrasal entailments using image similarities. Our assumption is that concepts expressed using different surface forms are mapped to similar visual representations, since humans tend to ground the meaning of phrases into the same visual perception irrespective of their culture and language.

Contact: Dan Han, AIRC-AIST, AIST Tokyo Waterfront, 2-3-26, Aomi, Koto-ku, Tokyo, 03-3599-8838, dan.han@aist.go.jp

2. Methodology

As we stated in Section 1., although the base RTE system uses lexical knowledge (i.e., WordNet) to recognize word entailments, it is still difficult to recognize semantic relations between phrases. One of the main reasons is that textual resouces on defining the relations among phrases are limited. Therefore, our effort in this paper focused on grounding phrases in image data, and comparing the phrases according to their visual representations. The comparisons of all the phrase pairs from a T-H pair release a set of visual features for the T-H pair, and by combining with the textual features that are obtained from the base RTE system, we obtain a feature vector for each T-H pair. So far, we have defined 10 visual features with two optional ones, and 9 textual feature for each T-H pair.

Visual Features

To generate phrase pairs between T and H, we used a tree mapping algorithm in [Martínez-Gómez and Miyao, 2016]. For each phrase pair, we first retrieved a set of images I_s for the source phrase (phrase from T) and a set of images I_t for the target phrase (phrase from H). Then, for $i_k^s \in I_s$ and $i_l^t \in I_t$, we obtain the vector representations $V(i_k^s)$ and $V(i_l^t)$ by using the first layers of a CNN.

We computer the cosine similarity as:

$$f(i_k^s, i_l^t) = cos(\boldsymbol{V}(i_k^s), \boldsymbol{V}(i_l^t)) = \frac{\boldsymbol{V}(i_k^s) \cdot \boldsymbol{V}(i_l^t)}{||\boldsymbol{V}(i_k^s)|| \cdot ||\boldsymbol{V}(i_l^t)||} \quad (1)$$

between every two images for a pair of phrases, which have the image sets, $I_s = \{i_1^s, i_2^s, ..., i_n^s\}$ and $I_t = \{i_1^t, i_2^t, ..., i_m^t\}$, to obtain the image similarity matrix:

$$f(I_s \times I_t) = \begin{bmatrix} f(i_1^s, i_1^t) & f(i_1^s, i_2^t) & \cdots & f(i_1^s, i_m^t) \\ f(i_2^s, i_1^t) & f(i_2^s, i_2^t) & \cdots & f(i_2^s, i_m^t) \\ \vdots & \vdots & \ddots & \vdots \\ f(i_n^s, i_1^t) & f(i_n^s, i_2^t) & \cdots & f(i_n^s, i_m^t) \end{bmatrix} \quad (2)$$

After we obtain the similarity matrix, we calculate four types of visual features, a) Max-Mean (the mean of all column max), b) Mean (the mean of the matrix elements), c) Max (the maximum value of the matrix), and d) Min (the minimum value of the matrix). We use these features to indicate how similar two phrases are, and for each T-H pair, we select two phrase pairs that have the highest and lowest Max-Mean scores, respectively. Thus, we employ the 4 visual features of the two phrase pairs as the visual features for the T-H pair.

Textual Features

From our observations, image groundings are ineffective in presence of negations, passive-active constructions, word-to-word verb relations (e.g. laughing and crying), antonym relations between any word in a phrase pair, and when comparing words that denote people of different gender (e.g. boy versus lady, man versus woman). Another important aspect is to understand why the logic prover produced an inconclusive judgment (*unknown*). To that end, we observe the state of the theorem proving: if the theorem proving stopped because a variable or meta-predicate[*1] failed to unify, then we consider this as a signal of logical conflict for which image grounding (or any type of paraphrasing) should be suppressed. Instead of considering the signals described above as rules to suppress the classifier, we simply add them as features to partially represent a T-H pair.

3. Experiment

We evaluate our system on the SemEval-2014 version of the SICK corpus [Marelli et al., 2014]. We split the corpus into three datasets: train $(4,500)$, trial (500), and test $(4,927)$. The distribution of the three entailment labels (yes/no/unknown) are $.29/.15/.56$. The average T and H sentence length was 10.6, were 3.6 to 3.8 words appeared in T and not in H or vice versa. We obtained 10 images for every phrase using Google Image Search API, and the image vector representations were obtained using the image miner and the feature extractor of [Kiela, 2016].

Our baseline is `ccg2lambda` [Martínez-Gómez et al., 2016][*2] when using only WordNet and VerbOcean to account for word-to-word lexical divergences. On the training data, `ccg2lambda` obtains an accuracy of 82.89%. Using our image-grounding classifier, we carried out 10 runs of a 10-fold cross-validation on the training data and the results are shown in Table 1. The results show that using image groundings to recognize phrasal entailments produce significant improvements in accuracy. Our preliminary experiment on training dataset obtained 1.06% higher accuracy (83.95 versus 82.89) with a standard deviation of 0.06% on 10 runs over the baseline. In this paper, we only report our results on training data.

System	Accuracy	Std.
ccg2lambda	82.89	–
ccg2lambda, `c+t`	76.60	0.03
ccg2lambda, `c+t+i(10)`	83.95	0.06

Table 1: Results (accuracy and standard deviation) of the classifier `c` on the training split of SICK dataset using text `t` and image `i` features for 10 images.

4. Conclusion

In this paper, we studied a method to compensate phrasal lexical divergences by grounding phrases in their visual representations. Although there have been works on multi-modal semantics, we believe that our work is the first successful attempt on aggregating visual features in a logic system to address the lack of perceptual grounding in an RTE system.

In the near future, we would like to extend our work in a few directions. One of them would be to investigate image features that account for the intersective meaning of adjectives and noun phrases. This could potentially signal hypernymy and other semantic subsuming relations. For instance, images of the concept "weapon" may contains images of the concept "sword". We would also like to import a distributional model to the visual features so that the system will learn from the enhanced influence of the critical images. Furthermore, it is also interesting to explore the interaction of other modalities such as the auditory perceptual information.

∗1 Logic predicate with no support from the surface form of the sentences, introduced by the semantic parsing.

∗2 https://github.com/mynlp/ccg2lambda

References

[Beltagy et al., 2013] Beltagy, I., Chau, C., Boleda, G., Garrette, D., Erk, K., and Mooney, R. (2013). Montague meets markov: Deep semantics with probabilistic logical form. In *Proc. of SEM*, pages 11–21.

[Kiela, 2016] Kiela, D. (2016). Mmfeat: A toolkit for extracting multi-modal features. In *Proc. of ACL System Demonstrations*, pages 55–60.

[Kiela and Bottou, 2014] Kiela, D. and Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proc. of EMNLP*, pages 36–45.

[Lai and Hockenmaier, 2014] Lai, A. and Hockenmaier, J. (2014). Illinois-LH: A denotational and distributional approach to semantics. In *Proc. of SemEval*, pages 329–334.

[Marelli et al., 2014] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proc. of LREC2014*, pages 216–223.

[Martínez-Gómez et al., 2016] Martínez-Gómez, P., Mineshima, K., Miyao, Y., and Bekki, D. (2016). ccg2lambda: A compositional semantics system. In *Proc. of ACL System Demonstrations*, pages 85–90.

[Martínez-Gómez and Miyao, 2016] Martínez-Gómez, P. and Miyao, Y. (2016). Rule extraction for tree-to-tree transducers by cost minimization. In *Proc. of EMNLP*, pages 12–22.

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