Improving Noun-Modifier Semantic Relation Prediction with Visual Information

Michael Haas <haas@cl.uni-heidelberg.de>

2013-09-03

Introduction

Base noun phrases consist of a head and a modifier. The head is a noun, the modifier may be an adverb, an adjective or a noun. A semantic relation holds between noun and modifier that further describes how the modifier modifies the head noun. Nastase and Szpakowicz [2003] investigate methods to predict these relations. They use a set of 50 semantic relations, which can be further grouped into five classes. The authors investigate two different noun phrase representations and three different machine learning approaches in order to predict the semantic relations. Additionally, they provide a data set consisting of 600 sense-disambiguated head-modifier pairs [1].

It is my goal to extend their approach with visual information. Recent work on semantics has gone beyond the text domain of traditional distributional representations of meaning and has become multimodal. Co-occurence information extracted from text corpora is unable to capture full word meaning as many facets of meaning simply go unmentioned. One reason is brevity: obvious facts that are common knowledge do not have to be mentioned. Other facts about concepts are not easily verbalized. Although we as humans can perceive and process taste, texture, smell and visual information well, talking about these aspects of the world is lossy.

The semantic classes consist of CAUSAL, PARTICIPANT, QUALITY, SPATIAL and TEMPORAL. CAUSAL include cause and effect relations, such as "flu virus". PARTICIPANT includes agent and object such as "student protest". SPATIAL relations indicate location, e.g. "home town". TEMPORAL relations place an occurrence in time, e.g. "morning coffee". QUALITY covers the remaining relations such as "material" in "brick house".

My hypothesis is that the addition of visual information improves classification performance. Furthermore, the improvement should be most pronounced in the QUALITY relation as the other relations are not easily expressed in pictures. For example, it is easier to obtain QUALITY information such as "brick house" from pictures than TEMPORAL or CAUSAL information.

Related Work

Leong and Mihalcea [2011] measure semantic relatedness between words and images. Taking synset definitions from WordNet and pictures from ImageNet, they use co-occurence information to model semantic relatedness as cosine similarity in vector space. Dimensions in vector space model individual synsets. Their feature extraction using SIFT descriptors [Lowe, 1999] and visual code words is essentially what inspired my approach.

Bruni et al. [2012] also integrate textual and visual information in a distributional framework. Their paper contributes the simple yet powerful idea that vector representations for text and images can be successfully concatenated to form a composite representation.

[1] http://www.site.uottawa.ca/∼vnastase/

Finally, my hypothesis is based on a long history of distributional models of meaning. The basic idea is that "statistical patterns of human word usage can be used to figure out what people mean", as noted by Turney et al. [2010] in their overview paper.

Data Representation & Classification

I build this implementation on the WordNet-based representation of base noun phrases given by [Nastase et al., 2006]. This representation relies on data annotated with WordNet synset IDs. [Fellbaum, 2010] Each word is represented by its hypernyms. Thus, the synset "true cat" is represented by the set ("true cat","feline", "carnivore", "placental",[..],"entity"). [Nastase et al., 2006] restrict the tree depth to 7. The entire set of hypernyms over the words in the data set, restricted to depth 7, forms the vocabulary/vector space. Every word is represented as a vector with 1 in the position corresponding to the hypernym. Every base noun phrase is the concatenation of a head vector and a modifier vector. The data set provided by [Nastase et al., 2006] is slightly adjusted as synset names have changed between WordNet 1.6 and 3.0. The vectors form the input for classifier.

ImageNet[Deng et al., 2009] links WordNet synsets to images. For every word in the data set, I download the set of image URLs from ImageNet and fetch a single random image. ImageNet is restricted to mostly nouns and not all nouns are annotated with images, therefore the 600 noun phrases from the original data set dwindle to 26 pairs[2]. To obtain more data, the algorithm ascends up to two levels in the hypernym hierarchy if the ImageNet server indicates that the originally requested synset does not have any images. The images are resized to 640px width to speed up processing.

A vector representation for the downloaded images is obtained with the OpenCV [Bradski, 2000] toolkit (version 2.4.6). Using the SIFT [Lowe, 1999] algorithm, I extract SIFT keypoints. The images are converted to grayscale beforehand as SIFT does not consider color information. These keypoints are converted into a vector representation of length 128, the SIFT descriptors. The vocabulary of 1000 visual code words is created by clustering all SIFT descriptors for all images. The process is illustrated in figure [Bruni et al., 2012]. Every image is then represented as a binary vector of length 1000 over this vocabulary. The vector indices are set to 1 if the corresponding visual code word is present, 0 otherwise.

The visually enriched noun phrase representation is then created by concatenating the wordnet-based vector with the image vectors obtained for head and modifier. In the end, I obtain 37 representations for base noun phrases.

Evaluation

The evaluation is set up as a binary classification task. For each of the five classes, a separate model is trained. For a given class, the classifier decides whether an instance belongs to the given class or not. The negative instances consist of the four other classes.

[2] A few more synsets are discarded due to faulty images and other implementation difficulties

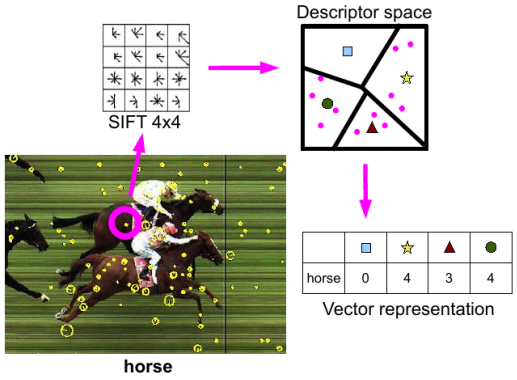

Figure 0.1: SIFT and the Bag of Visual Words model: From Keypoints to Descriptors to Visual Words Bruni et al. [2012]

Relation	IMG	WN	BL
CAUSAL	0.88	0.852	0.88
PARTICIPANT	0.587	0.650	0.411
QUALITY	0.757	0.697	0.411
SPATIAL	0.92	0.906	0.92
TEMPORAL	1	1	1

Table 0.1: F-Score for SVM and linear kernel

The evaluation closely follows [Nastase et al., 2006]. All experiments run ten-fold cross validation. The ML package is WEKA 3.7.9 [Hall et al., 2009] with LibSVM 3.17[Chang and Lin, 2011]. Nastase et al. [2006] use SVMLight with a linear kernel. I also choose a linear kernel as the default radial basis kernel does not outperform baseline. All other SVMLight settings remain default. I also report results obtained with the NaiveBayes classifier.

The baseline is the f-score obtained from the ZeroR classifier which assigns the most frequent class to all instances.

Results for the SVM classifier are reported in table 1. Results for NaiveBayes are shown in table 2. BL indicates the ZeroR baseline, WN is the basic WordNet model as per [Nastase et al., 2006] and IMG indicates the visually enriched model.

Results

The TEMPORAL data set only contains negative instances. The classifiers classify all examples correctly. SPATIAL and CAUSAL scores are close to baseline. With the SVM classifier, the IMG model performs slightly worse than baseline. NaiveBayes, on the other hand, slightly improves upon baseline by 0.012 points. Both PARTICIPANT and QUALITY beat the baseline by a large margin. For the SVM classifier, the WN model beats

Relation	IMG	WN	BL
CAUSAL	0.88	0.892	0.88
PARTICIPANT	0.731	0.622	0.411
QUALITY	0.807	0.659	0.411
SPATIAL	0.92	0.906	0.92
TEMPORAL	1	1	1

Table 0.2: F-Score for NaiveBayes

the IMG model on the PARTICIPANT relation. For NaiveBayes, IMG performs better. Best results are obtained with NaiveBayes on IMG for PARTICIPANT. QUALITY, for which I hypothesized that it would benefit most, has the biggest improvement (+0.396 over baseline, +0.148 over WN) with the NaiveBayes classifier. In both classifiers, IMG outperforms WN for the QUALITY relation.

Discussion

I successfully extend a linguistic model of semantic relations for base noun phrases with visual information. As expected, the QUALITY group of relations benefits most. The SVM classifier is outperformed by NaiveBayes, possibly due to poor parameter selection.

The model depends on sense-annotated data from WordNet and ImageNet. Possible future work includes an adaption to raw linguistic and pictorial data, where text is obtained from un-annotated corpora and pictures from possibly noisy sources such as image search engines. The current approach is limited to nouns only, as ImageNet does not provide images for other word classes. Furthermore, the feature extraction from pictures can be tuned more, such as the threshold in the SIFT feature detector. Finally, ImageNet provides multiple images for each synset. Using more than one image could improve performance, i.e. using the sum of the vector representations of multiple images. In the interest of reproducability and falsifiability, my code and the data files are available at https://github.com/mhaas/semrel-pictoral.

Bibliography

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics, 2012.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

Christiane Fellbaum. *WordNet*. Springer, 2010.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

Chee Wee Leong and Rada Mihalcea. Measuring the semantic relatedness between words and images. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 185–194, 2011.

David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

Vivi Nastase and Stan Szpakowicz. Exploring noun-modifier semantic relations. In *Fifth international workshop on computational semantics (IWCS-5)*, pages 285–301, 2003.

Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. Learning noun-modifier semantic relations with corpus-based and wordnet-based features. 2006.

Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.