

# Text Analysis for Automatic Image Annotation\*

Koen Deschacht

Marie-Francine Moens

*Interdisciplinary Centre for Law and ICT, Department of Computer Science, K.U.Leuven*

## 1 Introduction

In many situations texts and images in a document are highly associated, such as is the case of newspapers, (news-)websites, brochures and advertisements and catalogues. The text offers valuable information which could help to interpret the image, for example in the difficult task of automatic object recognition. In this paper we test the feasibility of automatically annotating images by extracting entities from the associated text which are likely to appear in the image. To determine the probability that an entity appears in the image, we take into account the text's discourse structure, the text's semantics and world knowledge that is not present in the text.

## 2 Entity classification

We classify all entities (i.e. nouns) in the text according to a semantic database. We use the WordNet [1] database, which organizes English nouns, verbs, adjectives and adverbs in synsets. A synset is a collection of words that have a close meaning and that represents an underlying concept. We combine two methods for entity classification. The first method assigns the correct synset to every noun in the text (i.e. to disambiguate the sense of the word) for which we use an efficient Word Sense Disambiguation system that was developed by the authors. This method does not offer a satisfactory solution for proper names, since the amount of proper names is possibly indefinite. Therefore, we use a second method in which we tag proper names using the Lingpipe Named Entity Recognizer. This package recognizes persons, locations and organizations.

## 3 Saliency

We discover which entities are in the focus of the text and which entities are only mentioned briefly, because we assume that more important entities in the text have a greater probability of appearing in the image. We define the *saliency* measure which represents the importance of an entity in a text. In earlier research we have developed a discourse segmentation module, which hierarchically and sequentially segments the discourse in different topics and subtopics resulting in a discourse tree, representing a hierarchical table of contents of the text [3]. The algorithm detects patterns of thematic progression in texts and can thus recognize the main topic of a sentence (i.e., about whom or what the sentence speaks) and the hierarchical and sequential relationships between individual topics. We compute for every entity a score (*Sali*) which is proportional to the depth of the entity in the discourse tree -hereby assuming that deeper in this tree more detailed topics of a text are described- and normalize this value.

Not all entities in the text appear in the discourse tree, and therefore we implement an additional refinement. The segmentation module already determines the main topic of a sentence. We can determine the relative importance of the other entities in a sentence by relying on the relationships between entities as

---

\*This is an abstract of the papers Deschacht, K. and Moens, M.-F. *Text Analysis for Automatic Image Annotation*, in proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, 2007 and Deschacht, K., Moens, M.-F. and Robeyns, W. *Cross-Media Entity Recognition in Nearly Parallel Visual and Textual Documents*, in proceedings of the 8th RIAO conference on Large-Scale Semantic Access to Content, Pittsburgh, 2007. The work reported in this paper is supported by the EU-IST project CLASS (IST-027978). We are grateful to Yves Gufflet from the INRIA research team (Grenoble, France) for collecting the Yahoo! News dataset.

signaled by the syntactic parse tree. For every entity we calculate the score  $Sal2$ , which is inversely proportional to the depth of the parse tree where the entity occurs, and the score  $Sal3$ , which is proportional to the number of nodes in the subtree corresponding to this entity. The salience is defined as the linear combination of these three scores ( $Sal1$ ,  $Sal2$  and  $Sal3$ ).

## 4 Visualness

The salience measure depends only on the thematic progression and syntactic structure of the text. Lacking from this measure is the fact that some entities never (or only indirectly) appear in an image, such as “thought” and “journey”. We therefore incorporate world-knowledge in a second measure, *visualness*, which is defined as the extent to which an entity can be perceived visually. We employ the similarity measure defined by Lin [2] which uses the hyponym/hypernym relation in the WordNet database to compute a semantic similarity between two synsets. This similarity measure gives a score between 0 and 1, close to 1 for synsets which are conceptually similar (e.g. “journey” and “voyage” with a similarity of 0.89) and close to 0 for synsets which are conceptually different (e.g. “journey” and “car” with a similarity of 0). We have manually selected 25 seed synsets in WordNet, where we tried to cover the wide range of topics we were likely to encounter in the test corpus, and set the visualness of these seed synsets to either 1 (visual) or 0 (not visual). We then define the visualness  $vis(s)$  of a given synset  $s$  as  $vis(s) = \sum_i vis(s_i) \frac{sim(s, s_i)}{C(s)}$  where  $s_i$  is a seed synsets and  $C(s)$  is a normalizing constant.

## 5 Results

We assume that the salience and visualness measures are independent for every entity. This allows us to compute the *probability of appearance* of an entity as the product of the salience and the visualness of the entity. We evaluate this measure on a corpus consisting of image-text pairs retrieved from the Yahoo! News website. Every picture has an associated text which discusses the event for which the picture was taken, with an average length of 40.98 words in which they discuss 15.04 entities, of which 2.58 appear in the image.

We perform a first evaluation on 100 image-text pairs where we predict, from analyzing the text only, what entities are visible in the associated image. One annotator has labeled every image-text pair with the entities (i.e. persons and other objects) that appear both in the image and in the text. A baseline system, which assumes that all entities in the text are visible in the image results in a precision of 15.62%, a recall of 91.08% and F-measure of 26.66%. Our system gives a probability of appearance for every entity. To simplify evaluation, we assume that all entities with a probability higher than a threshold of 0.4 are visible in the image. This results in a precision of 70.56%, a recall of 67.82% and a largely improved F-measure of 69.39%.

For the second evaluation we have annotated the entities that appear both in the image and in the text for 900 image-text pairs, and, for every image-text pair, sorted these entities on their perceived importance in the image. To compare this ranking with an automatic generated ranking, we use a penalty based evaluation measure where an entity on an incorrect position receives a penalty which is proportional with the difference with the position in the manual list. We compute a micro average for all 900 texts. A baseline approach where all entities are considered visible and ranked according to the position in the text results in a score of 56.98%. The ranking defined by the probability of appearance results in substantially improved score of 83.25%.

## References

- [1] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
- [2] Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proc. 15th International Conf. on Machine Learning*, 1998.
- [3] Marie-Francine Moens. Using Patterns of Thematic Progression for Building a Table of Content of a Text. *Journal of Natural Language Engineering*, 12(3):1–28, 2006.