Chapter 7

Learning Full-Sentence Co-Related Verb Argument Preferences from Web Corpora

Hiram Calvo
Nara Inst. of Science & Technology, Japan

Kentaro Inui
Tohoku University, Japan

Yuji Matsumoto
Nara Inst. of Science & Technology, Japan

ABSTRACT

Learning verb argument preferences has been approached as a verb and argument problem, or at most as a tri-nary relationship between subject, verb and object. However, the simultaneous correlation of all arguments in a sentence has not been explored thoroughly for sentence plausibility mensuration because of the increased number of potential combinations and data sparseness. In this work the authors present a review of some common methods for learning argument preferences beginning with the simplest case of considering binary co-relations, then they compare with tri-nary co-relations, and finally they consider all arguments. For this latter, the authors use an ensemble model for machine learning using discriminative and generative models, using co-occurrence features, and semantic features in different arrangements. They seek to answer questions about the number of optimal topics required for PLSI and LDA models, as well as the number of co-occurrences that should be required for improving performance. They explore the implications of using different ways of projecting co-relations, i.e., into a word space, or directly into a co-occurrence features space. The authors conducted tests using a pseudo-disambiguation task learning from large corpora extracted from Internet.

DOI: 10.4018/978-1-60960-881-1.ch007
1. INTRODUCTION

A sentence can be regarded as a verb with multiple arguments. The plausibility of each argument depends not only on the verb, but also on other arguments. Measuring the plausibility of verb arguments is needed in several tasks such as Semantic Role Labelling, since grouping verb arguments and measuring their plausibility increases performance, as shown by Merlo and Van Der Plas (2009) and Deschacht and Moens (2009). Metaphora recognition requires this information too, since we are able to know common usages of arguments, and an uncommon usage would suggest its presence, or a coherence mistake (v. gr. to drink the moon in a glass). Malapropism detection can use the measure of the plausibility of an argument to determine misuses of words (Bolshakov, 2005) as in hysteric center, instead of historic center; density has brought me to you; It looks like a tattoo subject; and Why you say that with ironing? Anaphora resolution consists on finding referenced objects, thus, requiring among other things, to have information about the plausibility of arguments at hand, i.e., what kind of fillers is more likely to satisfy the sentence’s constraints, such as in: The boy plays with it there, It eats grass, I drank it in a glass.

This problem can be seen as collecting a large database of semantic frames with detailed categories and examples that fit these categories. For this purpose, recent works take advantage of existing manually crafted resources such as WordNet, Wikipedia, FrameNet, VerbNet or PropBank. For example, Reisinger and Paşca (2009) annotate existing WordNet concepts with attributes, and extend is-a relations, based on Latent Dirichlet Allocation on Web documents and Wikipedia; Yamada et al. (2009) explore extracting hyponym relations from Wikipedia using pattern-based discovery and distributional similarity clustering. The problem with the semantic frames approach for this task is that semantic frames are too general. For example Anna Korhonen (2000) considers the verbs to fly, to sail and to slide as similar and finds a single subcategorization frame. On the other hand, n-gram based approaches are too particular, and even using a very big corpus (such as using the web as corpus) has two problems: some combinations are unavailable, or counts are biased by some syntactic constructions. For example, solving the PP attachment for extinguish fire with water using Google1 yields fire with water: 319,000 hits; extinguish with water: 32,100 hits. Resulting in the structure *(extinguish (fire with water)) instead of (extinguish (fire) with water). Thus, we need a way for smoothing these counts. This latter has been done by using Selectional Preferences since Resnik (1996) for verb to class preferences, and then generalized by Agirre and Martinez (2000) for verb class to noun class preferences. More recent work includes (McCarthy and Carroll, 2003), which disambiguate nouns, verbs and adjectives using automatically acquired selectional preferences as probability distributions over the WordNet noun hyponym hierarchy and evaluate with Senseval-2. However these aforementioned works have a common problem which is that they address separately each argument for a verb.

1.1. One Argument is Not Enough

Consider the following sentence:

There is hay at the farm. The cow eats it

We would like to connect it with hay, and not with farm. From selectional preferences we know that the object of eat should be something edible, so that we can say that hay is more edible than farm, solving this issue. From semantic frames, we have similar knowledge, but in a broader sense—there is an ingestor and an ingestible.

However, this information can be insufficient in some cases where the selectional preference depends on other arguments from the clause. For example:
Related Content

Opportunities and Challenges from Unlicensed Mobile Access (UMA) Technology  
[www.igi-global.com/chapter/opportunities-challenges-unlicensed-mobile-access/17525?camid=4v1a](www.igi-global.com/chapter/opportunities-challenges-unlicensed-mobile-access/17525?camid=4v1a)

Secret Sharing with k-Dimensional Access Structure  
[www.igi-global.com/chapter/secret-sharing-dimensional-access-structure/21314?camid=4v1a](www.igi-global.com/chapter/secret-sharing-dimensional-access-structure/21314?camid=4v1a)

Predicting Key Recognition Difficulty in Music Using Statistical Learning Techniques  
Ching-Hua Chuan and Aleksey Charapko (2014).  *International Journal of Multimedia Data Engineering and Management* (pp. 54-69).  
[www.igi-global.com/article/predicting-key-recognition-difficulty-in-music-using-statistical-learning-techniques/113307?camid=4v1a](www.igi-global.com/article/predicting-key-recognition-difficulty-in-music-using-statistical-learning-techniques/113307?camid=4v1a)

Reducing Processing Demands for Multi-Rate Video Encoding: Implementation and Evaluation  
[www.igi-global.com/article/reducing-processing-demands-multi-rate/69518?camid=4v1a](www.igi-global.com/article/reducing-processing-demands-multi-rate/69518?camid=4v1a)