Spatial Text Labelling and Discovering Spatial Ontologies

Under Supervision Of
Prof. Amitabha Mukherjee

PRACHI GUPTA (12485)
SATYAM KUMAR SHIVAM (11917658)
Indian Institute of Technology, Kanpur

October 4, 2015

1   Problem Statement

The objective of our project is to identify and label spatial keywords in the given corpus and find spatial ontologies in these keywords. Instead of using any parser based method, our approach is based on word vector models. The aim is to find relationships among spatial keywords taken from a corpus in language other than English (eg Hindi).

2   Motivation

Spatial Role Labelling is useful in NLP to perform tasks like direction interpretation, verbal spatial description and metric data integration etc. A lot of work has been done earlier in this field but most of those works use a grammar based parser to find relationships among different objects. So, it becomes difficult to extend them to other languages. Hence, we try to use Word2Vec to obtain these relationships. If successful, we'll further try to extend the work to Hindi.

3   Approach

1. **Spatial Text Labelling**

   - Static Spatial Relation: The following annotations form a tuple of static spatial relations:
     Trajector: denotes a central object of a spatial scene.
     Landmark: denotes a secondary object of a spatial scene.
     Spatial Indicator: signals a spatial relation between objects.
     Spatial Relation: relation which holds between spatial markables in text.

   - Dynamic Spatial Relation: The following roles are given:
     Trajector: denotes an object which is in motion.
     Motion Indicator: signals the motion of trajector along a path.
     Path: denotes the path of the motion of the trajector.
     Landmark: capture the spatial context of the motion.
     Distance: denotes an absolute or relative distance of a motion.
     Direction: denotes an absolute or relative direction of a motion.
     Spatial Relation: relation which holds between spatial roles.

2. **Word Vector based Modelling**
   We train Google's Word2Vec model based on the spatially annotated text. The model will separate clusters for the above mentioned spatial labels.

4   Data Set

The dataset SpRL for the Spatial Role Labelling shared task at SemEval-2013 would be taken.

1. IAPR TC-12: This contains data for static (no motion involved) spatial relations. For example: a lake in the forest

2. ConfluenceProject: This has data for dynamic (involving motion) spatial relations. For example: I stepped into the small forest and followed down a dried creek

For Hindi, we'll build our dataset using articles taken from different websites like www.jagranyatra.com and neerajjaatji.blogspot.in

References

[1] Kolomiyets, Oleksandr, et al. "Semeval-2013 task 3: Spatial role labeling." Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013). 2013.

[2] Kordjamshidi, Parisa, and Marie-Francine Moens. "Global machine learning for spatial ontology population." Web Semantics: Science, Services and Agents on the World Wide Web 30 (2015): 3-21.

[3] Pustejovsky, James, Jessica Moszkowicz, and Marc Verhagen. "A linguistically grounded annotation language for spatial information." (2013).

[4] Clough, Paul. "Extracting metadata for spatially-aware information retrieval on the internet." Proceedings of the 2005 workshop on Geographic information retrieval. ACM, 2005.

[5] "Kolomiyets/sprl2013." GitHub. N.p., n.d. Web data repository.