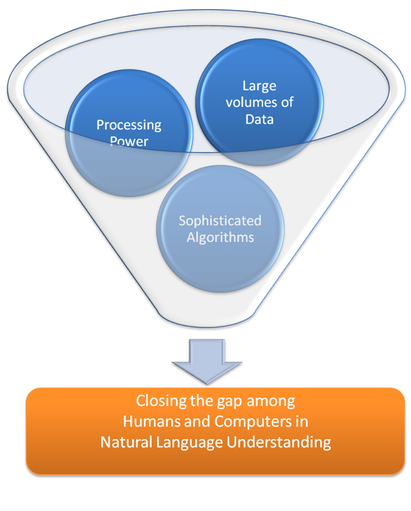
Contribution to research new models of knowledge extraction on BigData systems

Héctor Cerezo-Costas, Advisor: F.Javier González-Castaño[1]

[1]Department of Telematics Engineering, University of Vigo

Motivation



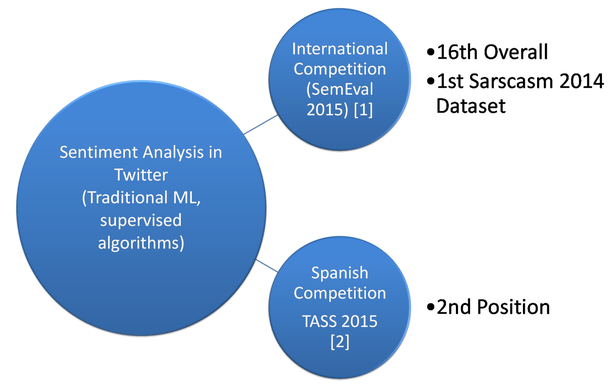
Thesis Objectives

Objective 1
- Fast context adaptation of NLP models

Objective 2
- Unsupervised algorithms with application to NLP

Objective 3
- Improve language understanding with Deep Learning technologies

Previous Work



Results & Discussion

Learning Semantic Sentence Representations

Useful for finding concepts with fuzzy searches. Application in professional writing environments.

Problems:

Infinite Possibilities

"an animal that is commonly kept as a pet and is famous for loyalty to humans"

Fuzzy Definitions

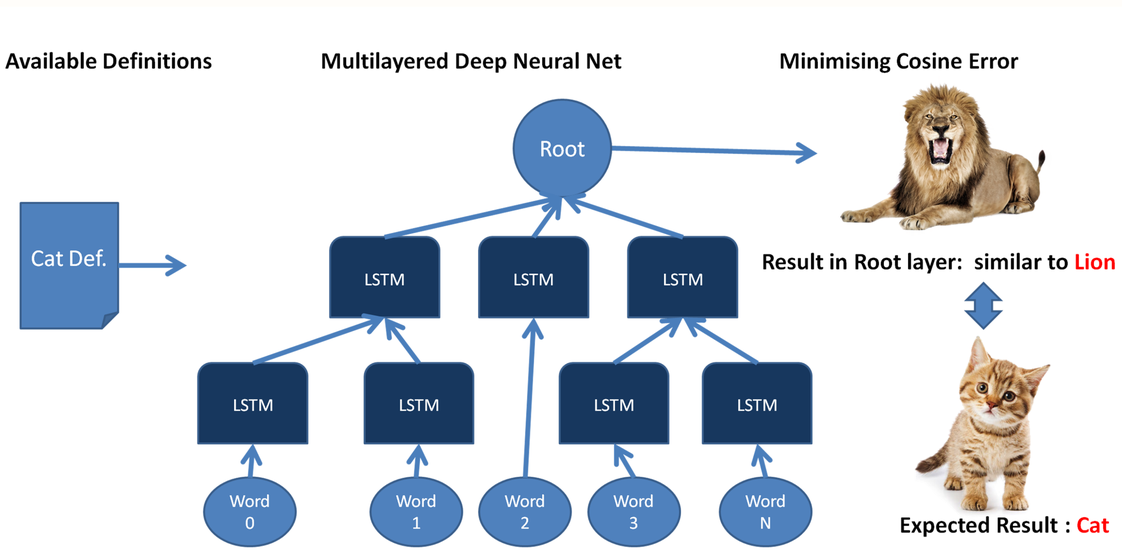
"something that is a circle but also flat"

Reasoning Requirements

"one of the directions on a compass that points right when you look at it"

Approach: Reverse Dictionaries

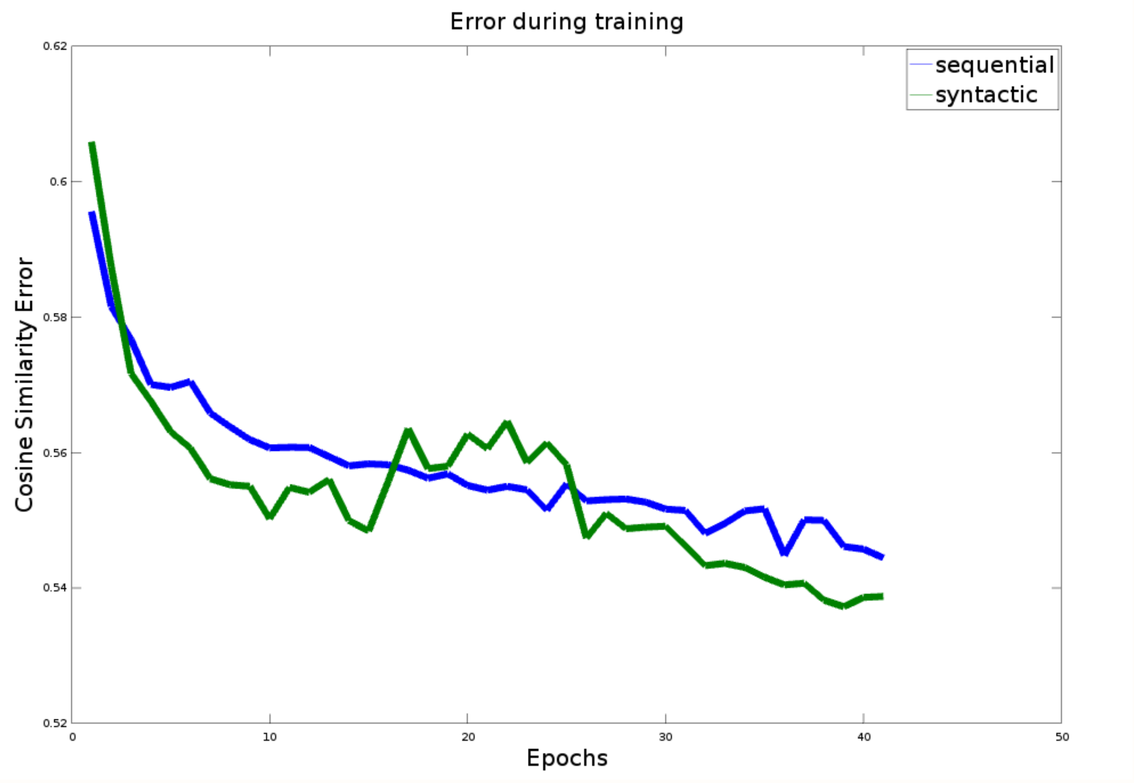
- Given a definition try to guess the most suitable word from a large collection of possible targets (e.g. 80000 words available in a dictionary)



- Semi-supervised Strategy. Models trained with dictionary definitions and a subset of Wikipedia descriptions
- Model: Deep Learning Tree LSTM Network [3]
- Error measurement: cosine similarity (target words vs root layer)

Assumption: Using syntactic joints in the deep engine will produce better results than using sequential joints (most common approach)

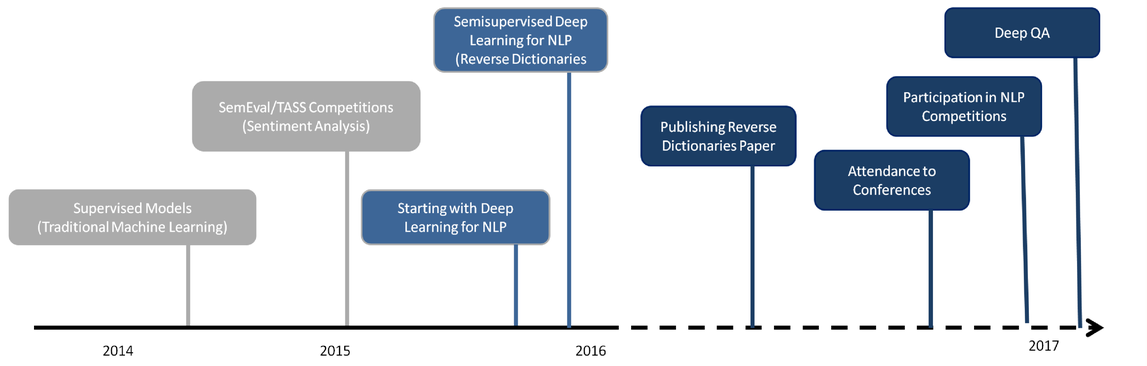
Preliminary Results



The graph shows the evolution of the avg. error in the test records as a function of the number of epochs.

Work in Progress: functional tests (e.g. avg. ranking of target words, accuracy discerning target words, etc), different network configurations,etc.

Research Plan (Next Year)



References

[1] H. Cerezo-Costas, and D. Celix-Salgado, (2015, June). Gradiant-analytics: Training polarity shifters with CRFs for message level polarity detection. *In Proc. of the 9th Int. Workshop on Semantic Evaluation (SemEval 2015) (pp. 539-544).*

[2] T. Alvarez-López, J. Juncal-Martınez, M. Fernández-Gavilanes, E. Costa-Montenegro, F.G. González-Castano, H. Cerezo-Costas, and D. Celix-Salgado GTI-Gradiant: A Hybrid Approach for Sentiment Analysis in Twitter. *In Proc. of TASS 2015*

[3] K. S. Tai, R. Socher and C.D. Manning, (2015). Improved semantic representations from tree-structured long short-term memory networks.

Workshop on Monitoring PhD Student Progress, 13 June 2016, Vigo, Spain