NERTUW : *Named Entity Recognition on tweets using Wikipedia*

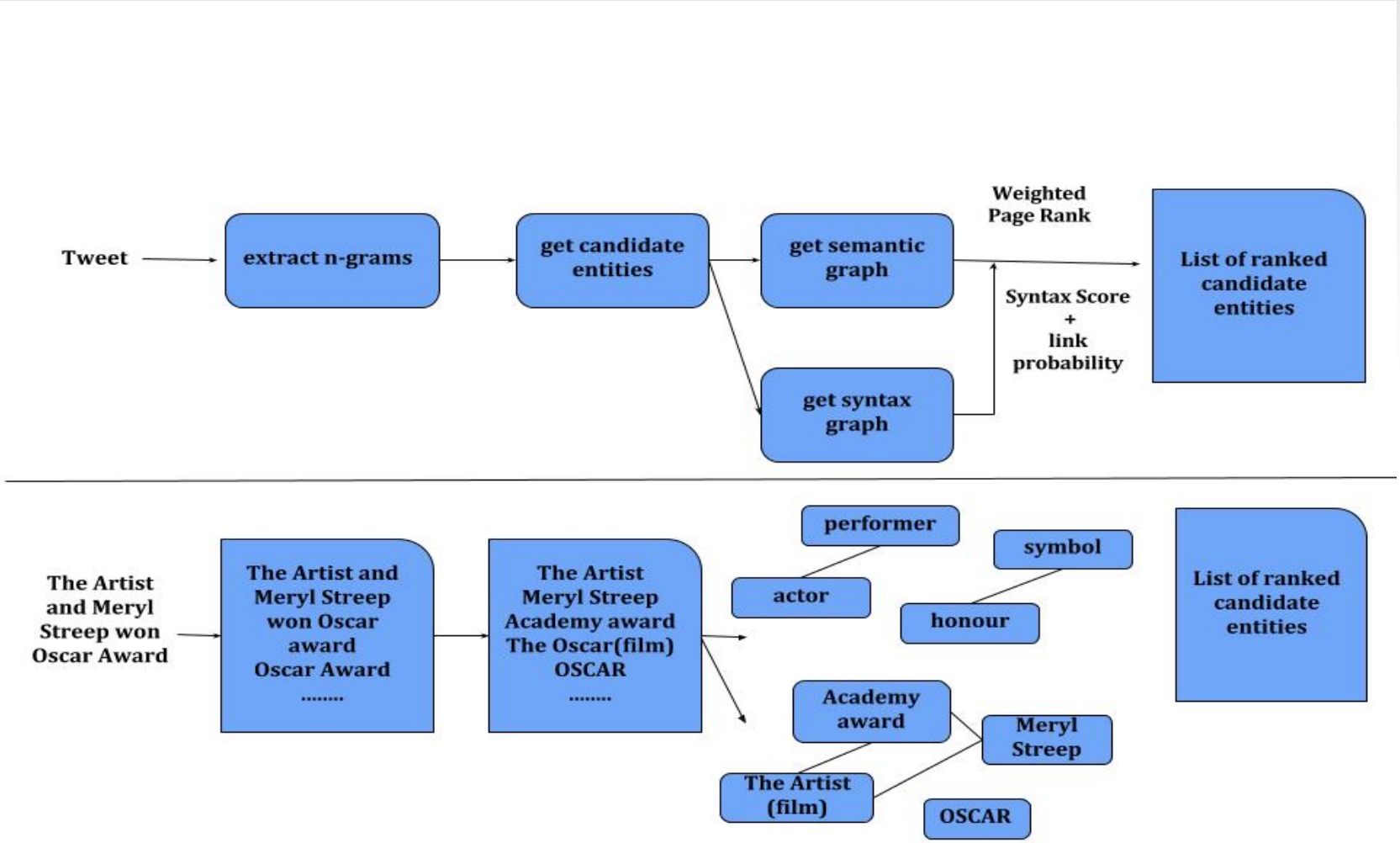
CONTRIBUTIONS

- An approach which utilizes the titles, anchors and infoboxes contained in Wikipedia and a little information from Wordnet and the context information in tweets to recognize, disambiguate and classify named entities in tweets
- Does not require any training data ie, no human labelling effort needed.
- Uses a Word sense disambiguation approach to disambiguate the context which in turn is used to disambiguate the named entities in tweets.

APPROACH

- Tweet is split into ngrams, those with a low link probability are pruned.
- Each ngram is matched lexically with Wikipedia article titles to get the candidate entities
- Syntax analyser generates a syntax graph using YAGO's type relation and the context information in the tweet. Each vertex represents a Wordnet synset and the edges represent a relation between them in Wordnet.
- Semantic analyser creates a semantic graph with candidate entities as vertices and edges based on their similarity score.
- Page Rank is applied on the syntax graph and the scores of vertices are used as prior for a Weighted Page Rank applied on the semantic graph.

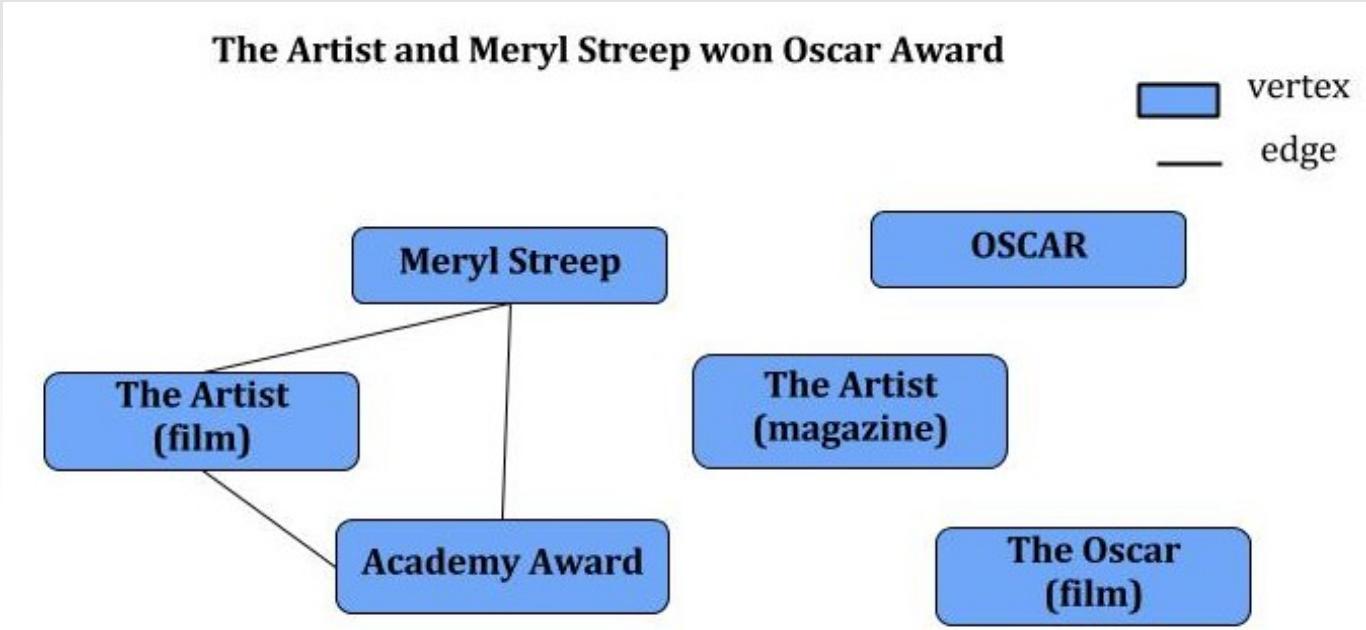
APPROACH


Tweet
extract n-grams
get candidate entities
get semantic graph
get syntax graph
Weighted Page Rank
Syntax Score + link probability
List of ranked candidate entities
The Artist and Meryl Streep won Oscar Award
The Artist and Meryl Streep won Oscar award Oscar Award .....
The Artist Meryl Streep Academy award The Oscar(film) OSCAR .....
performer
actor
symbol
honour
Academy award
Meryl Streep
The Artist (film)
OSCAR
List of ranked candidate entities

Syntax Graph

- Each vertex represents a Wordnet synset obtained from the context of the tweet and Yago.
- Each edge represents a relation between the synsets in Wordnet
- Page Rank Algorithm is applied on this graph to get syntax scores for each candidate entity

Semantic graph



- Each vertex represents a candidate entity
- An edge is added between the vertices if the similarity between them is higher than a threshold.
- Weighted Page Rank algorithm is executed on this graph with the syntax scores as priors to obtain the final ranking of entities.

Entity Classification

- Each ngram which has a candidate entity in the semantic graph is considered as a named entity.
- For each ngram, the candidate entity with the highest page rank in the semantic graph is given to a named entity classifier.
- Named entity classifier uses keywords in the infobox of the Wikipedia pages of the candidate entity to classify it as person, location, organization or miscellaneous.
- Unique keywords with maximum occurence, pertaining to each entity type provided in the training data were extracted to classify the named entities.

Thank you!