Aspect based Sentiment Analysis

Enayat Ullah[1];Ankit Singh[2];
[1]Department of Mathematics and Scientific Computing; [2]Department of Computer Science & Engineering
Supervisor: Prof. Amitabha Mukerjee

ABSTRACT

Sentiment Analysis is widely used to adjudge the semantic orientation of a text unit. However, a major challenge in sentiment analysis is the identification of the entities. the polarity is attributed to.

Aspect is an explicit reference of an entity towards which an opinion is expressed.

For example: The food of restaurant is amazing.
Aspect: food;
Polarity: Positive

Aspect-bases Sentiment Analysis is a two-fold SemEval[1] task, wherein first the aspect term is identified from the sentence and then polarity of the opinion corresponding to that aspect is adjudged.

A linear-chain CRF is trained with features based on word vectors and text processing techniques(POS, dependency parse) to sequentially label the aspect term in a sentence. A Maximum Entropy classier then identifies the polarity corresponding to the aspect. with features based on cosine similarity with words rom sentiwordnet.

INTRODUCTION

Sentiment analysis refers to identification and extraction of subjective impressions from text sources. It aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. In general, a binary composition of opinions is assumed: for/against, like/dislike, good/bad etc. However, sometimes an opinion can also be categorized into a neutral sentiment, if the polarity of the observed opinion fails to exceed a certain threshold. In such cases, we have a triplet of semantic orientations possible.

Sentiment analysis finds it's application in various disciplines; in Information Extraction, it is used to discard subjective information, in Question-Answering, it identifies opinion-oriented questions; in news sources, detecting if there is bias expressed by the author.

Various approaches have been put to use to identify aspects from sentences. Bing Lui et al. used frequency of noun phrases, followed by a redundancy pruning to identity the feature corresponding to a review[2]. Yejin Choi et al. performed semantic tagging using conditional random fields with features based on Capitalization, syntactic chunking to extract sources of opinions from texts[3].

FEATURES

MaxEnt Features:
- Nearest Adjective and it's polarity
- Aspect term dependent on an opinion word?
- Minimum Cosine similarity between the adjective and words from sentiwordnet[7].

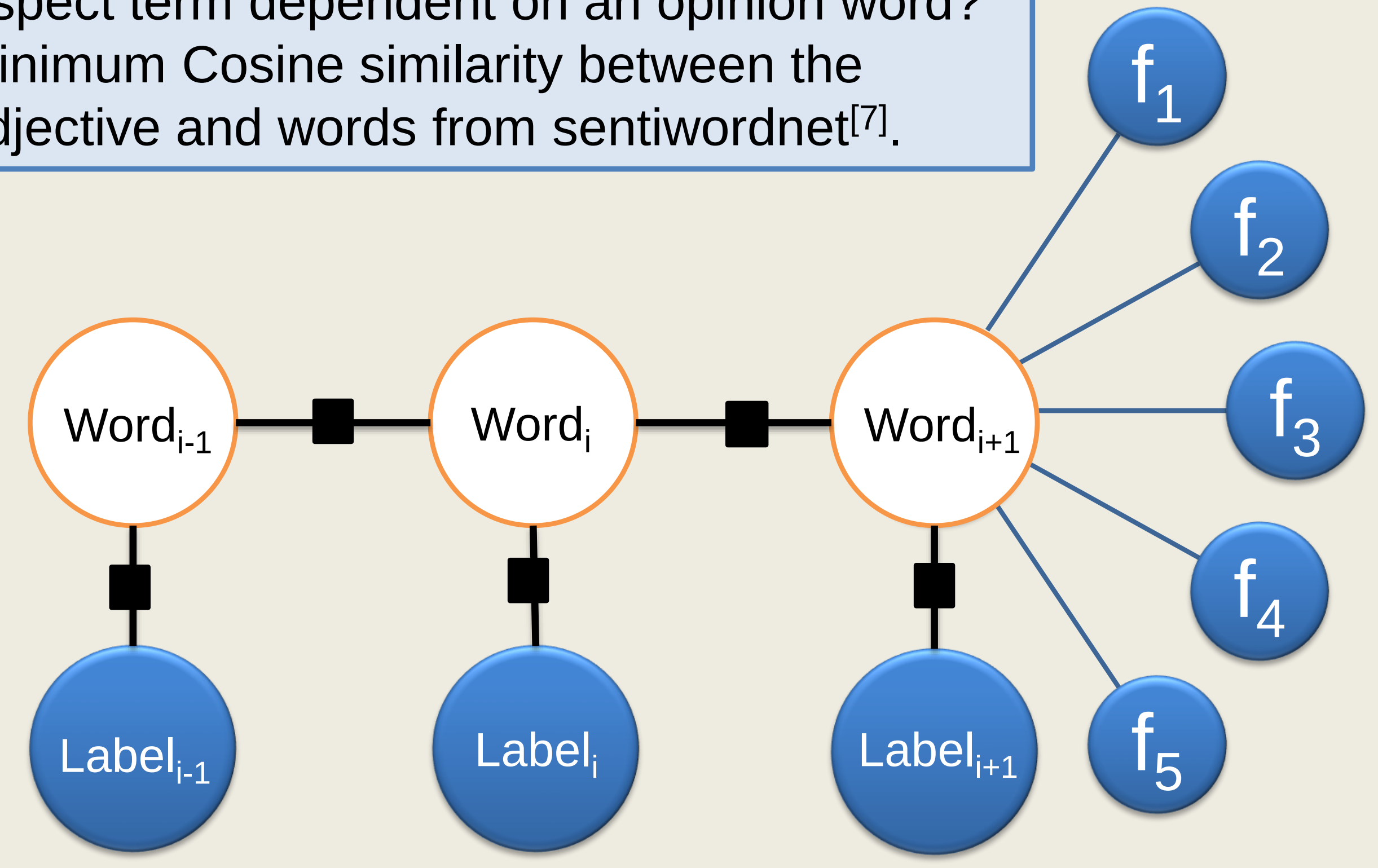


Figure 1. **Linear chain CRF with features**

f_1	f_2	f_3	f_4	f_5
Cosine Similarity with domain Centroid	POS tag of word	Word dependency on opinion word	N-gram words	Capitalization of word

Table 1. **Table of Features**

IMPLEMENTATION

Preprocessing: A word2vec[4] model is trained on 11.7 GB dump of English Wikipedia Corpus to obtain word vector representations each of dimensionality 100. The training data provided my SemEval consists 0f 1386 sentences, each tagged with one or more aspect terms and the corresponding polarity in an XML file.

The XML is parsed and punctuations are removed from the sentences(excluding –') . Stanford corenlp library is used to POS tags, word tokens and dependency parse of sentences[5].

Aspect term Identification:
CRFs are a type of discriminative undirected probabilistic graphical model used to encode known relationships between observations and construct consistent interpretations[6].

The formula below defines the linear-chain CRF: $y = \{yt\}_{t=1}^{T}$ $x = \{x_t\}_{t=1}^{T}$ are label sequence and observation sequence respectively, and there are *K* arbitrary feature functions $\{fk\}_{1 \le k \le K}$ and the corresponding weight parameters$\{\theta_k\}_{1 \le k \le K}$.

$$\mathrm{P}(\mathrm{y}|x) = \frac{1}{Z(x)} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(yt, y_{t_1}, x, t))$$

Polarity Detection:
A Maximum Entropy model defines the conditional distribution of the class *(y)* given an observation vector *x* where θ_k is a weight parameter to be estimated for the corresponding feature function $f_k(x, y)$

$$\mathrm{P}(\mathrm{y}|x) = \frac{1}{Z(x)} \exp(\sum_{k=1}^{K} \theta_k f_k(x, y))$$

Z(x) is a normalizing factor over all classes to ensure
a proper probability
The feature functions of CRF and maxEnt is provided in Fig. 2

RESULTS

Domain	Precision	Recall	Accuracy	F1
Laptop	0.5254	0.6823	0.9132	0.5936
Restaurant	0.5769	0.7443	0.9429	0.6522

Table 2. **Sub-Task 1: Aspect Term Identification**

Polarity	Domain	Precision	Recall	Accuracy	F1
Positive	Laptop	0.6142	0.6731	0.7832	0.6132
	Restaurant	0.6433	0.6876	0.7656	0.6432
Negative	Laptop	0.5457	0.7033	0.7832	0.5997
	Restaurant	0.5212	0.6746	0.7656	0.5587

Table 3. **Sub-Task 2: Polarity Detection**

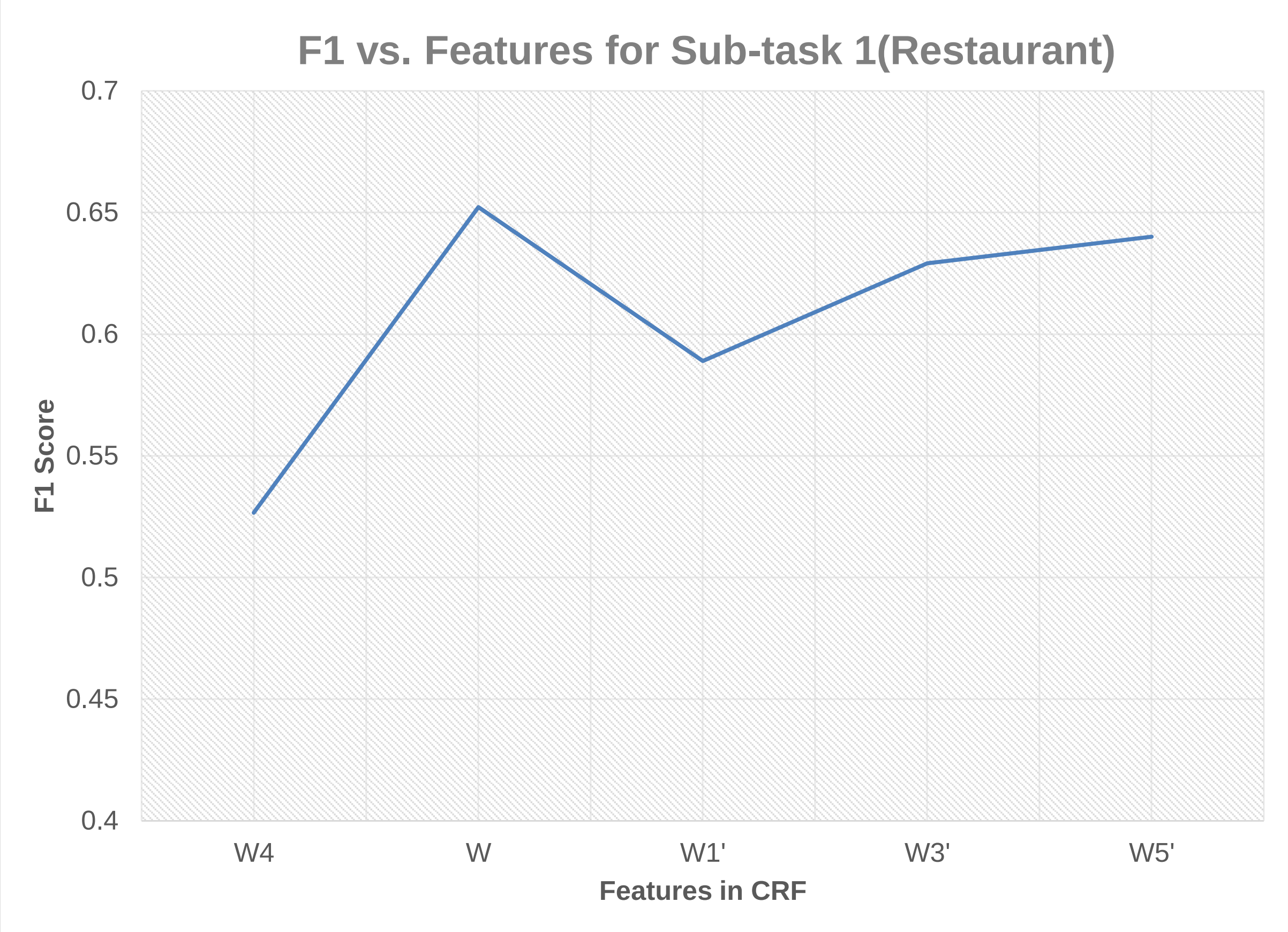
DISCUSSION

The trained CRF and maxEnt is tests on test data provided by SemEval. The test data XML file contains 787 sentences , each tagged with one or more aspect terms and the corresponding polarity.

Evaluation scores of a baseline algorithm(SVM with linear kernel) provided by SemEval is summed below:

Domain	Task	Score
Laptop	Aspect term Extraction	F-1: 0.3858
Laptop	Polarity Detection	Accuracy: 0.7647
Restaurant	Aspect term Extraction	F-1: 0.4868
Restaurant	Polarity Detection	Accuracy: 0.7174

Following is a graph depicting the variation of F-1 scores with the different features taken for CRF for Restaurant domain(W4: With only feature f_4; W1': Without feature 1)



REFERENCES

[1]. SemEval-2015 Task 12: Aspect Based Sentiment Analysis

[2]. M. Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews

[3]. Yejin Choi and Claire Cardi, Ellen Riloff and Siddharth Patwardhan, Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns.

[4]. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean; Efficient Estimation of Word Representations in Vector Space.

[5]. The Stanford Natural Language Processing Group, Stanford CoreNLP.

[6]. John Lafferty, Andrew McCallum, Fernando C.N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

[7]. A Esuli, F Sebastiani; Sentiwordnet: A publicly available lexical resource for opinion mining.