1. Aim
- To find aspects, and their opinion carrying descriptors, within sentences of reviews using a set of syntactic rules.
- A clustering algorithm for identifying and clustering similar aspects, using similarity features based on context and knowledge based similarity.

2. Motivation
- Online reviews are an important online resource. However going through a large number of reviews is a tedious task.
- Need to get opinions about a specific aspect from all reviews automatically.
- Need to get reviews related to all related aspects without specifying explicitly.

3. The Data
- Dataset consists of online reviews of businesses provided by Yelp for “Yelp Dataset Challenge 2014” consisting of 1,125,458 reviews of different businesses.
- The reviews of one hotel related business was taken as the final dataset, as it contained a lot of diverse aspects which could be identified.
- 6,784 review sentences were used in clustering.

1. Extraction of Aspect- Descriptor pairs
Table below shows the custom syntactic rules we have created to identify and extract aspects and their descriptors.

<table>
<thead>
<tr>
<th>Chunk labels</th>
<th>Rule</th>
<th>Pair extracted</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[a-zA-Z]+[a-zA-Z]+</td>
<td>[NP][I][B][A]</td>
<td>They have broken windows</td>
</tr>
<tr>
<td>B</td>
<td>[a-zA-Z]+[a-zA-Z]+</td>
<td>[NP][I][B][A]</td>
<td>[NP][I][B]</td>
</tr>
<tr>
<td>C</td>
<td>[NP][I][B][A]</td>
<td>[NP][I][B]</td>
<td>Opening is always hectic.</td>
</tr>
<tr>
<td>D</td>
<td>[NP][I][B][A]</td>
<td>[NP][I][B]</td>
<td>Not cooler is amazing.</td>
</tr>
<tr>
<td>E</td>
<td>[NP][I][B][A]</td>
<td>[NP][I][B]</td>
<td>Rooms are clean.</td>
</tr>
<tr>
<td>F</td>
<td>[NP][I][B][A]</td>
<td>[NP][I][B]</td>
<td>Ingredients are great for the money.</td>
</tr>
<tr>
<td>G</td>
<td>[NP][I][B][A]</td>
<td>[NP][I][B]</td>
<td>[NP][I][B]</td>
</tr>
</tbody>
</table>

2. Clustering of Related Aspects
Step 1: Connect aspects containing similar terms.
Aspects which are exactly similar or are almost exactly similar in case of multigrams are aggregated into a list of similar aspects, or an aspect-set. Aspects which share some words are likely to belong to the same cluster, for example “pool table” and “wooden pool table” most likely refer to the same aspect.

Step 2: Calculation of feature values for clustering.
3 types of similarity features calculated:
- Context or Co-occurrence of aspects
- External or Knowledge based similarity
- Distributional similarity of descriptors

Step 3: Clustering step.
Sort by similarity values and check if overshoot(TW)
Aspects which satisfy the threshold value enter in Ah smallest
Ah already in another smaller
Ah and its similar numbers are added to Ah with greater similarity

Definitions and Background
- **Aspects** are small units of a business which can be a product, an abstract entity like service, décor etc.
- **Descriptors** express opinions and sentiments about an aspect, which can be further used in generation of summaries for the aspects. Adjectives, Participles, Articles, Possessive Pronouns and Prepositional Phrases can describe or modify a noun.
- **Aspect-descriptor pair** - An aspect-descriptor pair consists of an aspect and the descriptor of that aspect. e.g. (sandwich, tasty) in “This is a tasty sandwich”.

Conclusions
- In this paper, we study the problem of aspect discovery and clustering.
- We first discovered aspect-descriptor pairs from reviews. Then we proposed three features and metrics for aspect similarity and an unsupervised clustering method
- The aspect similarity features proposed performed worked well with both clustering algorithms and have proven to be better than baseline method.
- The experiments are preliminary, and our method has yet to be tested on different datasets and domains. More parts of speech like verbs can be considered as descriptor words.
- In future, we plan to generate natural language summaries of aspect clusters, to highlight the constituent aspects and their descriptors in a meaningful manner.

References