Building an English-Chinese advertisement corpus

Scarlet W. Y. Li, Shan Wang, Grégoire Winterstein
The Hong Kong Institute of Education
BACKGROUND, MOTIVATION
Background

• The language of advertisement has been studied rather extensively (since Leech, 1966)

• However:
  – Most studies are qualitative
  – Most studies focus on one language (some exceptions: Tanaka, 1994)
  – Beyond a discourse analysis approach, the study of advertisement also offers interesting insight for semantics and pragmatics
Goals

• Construction of a **bilingual advertisement** corpus:
  – Chinese (Mandarin/Cantonese) and English

• **Annotation** of the corpus
  – Argumentative relations
  – Alignment of discourse markers

• Open **access** of the data
Argumentation theory

• Linguistic Argumentation Theory (Anscombre & Ducrot, 1983) postulates that every utterance targets an argumentative goal

• At its core, LAT studies argumentative markers and how they affect the argumentative potential of an utterance
  – John was barely late. ⇒ John is reliable/serious.
  – John was almost late. ⇒ John is not reliable/serious

• Markers have received detailed formal descriptions (Anscombre & Ducrot, 1983; Winterstein, 2010), but with little empirical backing
Argumentation in Advertisement

• A recurring problem when studying argumentation is the **abduction problem**:  
  – Given an utterance, how is it possible to reconstruct the goal targeted by the utterance?

• Generally, the question cannot be answered from linguistic material alone, which makes massive quantitative approaches impractical

• Advertisements have the advantage of having a relatively **clear/obvious goal**: promotion of a service/sell a product etc.
Argumentative Markers

![Table 1. Types of marker in the corpus](image)

<table>
<thead>
<tr>
<th>Marker(s)</th>
<th>Valence 1</th>
<th>Valence 2</th>
<th>Valence 1 or 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almost, (but) also, exactly, indeed, just, merely, moreover, nearly, (not) only, probably, quite, really, totally, very, even if</td>
<td>But, yet, because (of), since, though, unless, however, despite, in addition</td>
<td>Even (though)</td>
<td></td>
</tr>
</tbody>
</table>

- Examples in the corpus:
  - Return Fare from just HK$4,850
  - Our schools' international curriculum uses English as the language of instruction. **However**, Chinese also plays an important part in the curriculum, as all students are required to learn Putonghua, the official language in China.
METHODOLOGY
Methodology

• **Manual collection** of material taken from:
  – Internet
  – TV advertisements

• All material is **bilingual** (either Written Chinese / English or Cantonese/English)
  – The same content exists in both languages
  – Most of the material was prepared for the HK market

• **Manual annotation** of
  – Argumentative information
  – Alignment information between languages
Metadata

- Advertisement and promotional material in both English and Chinese used by Hong Kong based companies.

- Two main sources of material:
  - Texts from the official promotional websites of various companies (1255 texts)
  - Transcripts of TV advertisements (150 ads)
Metadata

• Metadata descriptors for the Advertisements:
  – The name of the advertising company
  – The nature of its services
  – A link to the website/ TV ad (if available online)
  – The type of advertised product
  – A screen capture in the case of a website (not used at the moment)
Stimulate all 5 senses at once
All aboard the magical spaceship for an exciting ride around breathtaking Lantau Island.
Fly over the Big Buddha, before diving underwater with Hong Kong’s famous white dolphins.
Indulge all your senses in the magnificence of Lantau, sky, land and sea, from the highest mountain peak, to the deepest gorge.
only $245 up
Online offer 10% off
Ngong Ping Walk n Motion Pass

五维动感影院 激新感官体验
踏上神秘飞船，与船长展开首次刺激紧张的飞行任务！
飞越天坛大佛，潜入中華水族秘境，穿梭心经密林，带你从多角度遨遊大嶼山，感受不一样的视觉、听觉、嗅觉、触觉及动感全方位震撼。
只需$245起
网上尊享9折优惠
昂坪大动感同行套票
Argumentative annotation

• Annotation done in two steps:
  – **Automatic** annotation of argumentative markers
  – **Manual** annotation of scope and bilingual relations

• Two phases
  – English / Chinese (done)
  – Chinese / English (underway)

• Annotation tool: **Webanno** (Yimam et al., 2013)
Manual annotation

• For all the markers automatically pre-annotated:
  – Annotation of the **scope** of the marker
  – **Link** between scope and marker
    
    *John almost hit the wall.*
  
  – **Alignment** with a marker in the other language
Bilingual relations

• Bilingual relations were annotated between:
  – Argumentative Markers
  – Scope of the markers
• Use of the scheme of Bond & Wang (2014):
  – Synonym (=): 因為/because
  – Pragmatic Link (≈): 咁/but
  – Lexical Link (~): 更可/also
  – Partial translation (#):
    *China 's taxation can be categorized/稅收劃分為*
  – Hypernym (>)
  – Hyponym (<)
  – Antonym (!)
# More examples

<table>
<thead>
<tr>
<th>Relations</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Synonym: =         | We **also** sell examination publications on behalf of our partnering examination bodies.  
                      | 我們**亦**有為合作機構代售考試刊物。                                    |
| **Partial**        | It is **also** used in some Light Buses, Vans and Passenger Cars.        
                      | 電裝在香港擁有領導地位，其中超過百分之九十五的雙層及單層巴士都裝用電裝空調系統，另多款私家車、輕型客貨車及小巴**亦有**採用。 |
| **Translation:**   |                                                                         |
| **Pragmatic Link:**| This may be due to large amounts of cash being excessively invested in fixed assets, or **because** the inventory turnover ratio is low, or credit policy is too loose, etc.   
                      | 其中的原因可能是大量現金被過多地投放於固定資產，也**可能**是存貨周轉率低，或信用政策過於寬鬆等。 |
An example

- [...] which **not only** (=) resolve the problem (<) to facilitate the business development for enterprise, **but also** (=) **effectively use various finance tools** (=) to raise capital for the enterprise to facilitate their business development.

- 不僅為企業解決了資金鍊的問題，更有效地利用各種融資工具為企業籌集發展業務的資金。
Contents of the corpus

• **1405** documents in total
  – **1255** texts from internet
  – **150** TV ads transcripts
• **150** different companies
• Varied services:
  – Banking, finance
  – Entertainment
  – Retail
  – Food industry
  – ...
## Contents: sizes

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Tokens</td>
<td># Char</td>
</tr>
<tr>
<td>Web material</td>
<td>152090</td>
<td>301254</td>
</tr>
<tr>
<td>TV advertisement</td>
<td>13026</td>
<td>21680</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>165116</strong></td>
<td><strong>322934</strong></td>
</tr>
<tr>
<td>Web material</td>
<td>121.2</td>
<td>240.0</td>
</tr>
<tr>
<td>TV advertisement</td>
<td>86.8</td>
<td>144.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>117.5</strong></td>
<td><strong>229.8</strong></td>
</tr>
</tbody>
</table>
EXAMPLES OF USE
Use of the corpus

• The corpus has different practical uses:
  – Study of the advertisement discourse in a comparative perspective
  – Study of argumentative markers and their crosslinguistic differences
  – Use for tasks related to opinion mining (Pang & Lee, 2008), and more generally machine-learning related tasks
Proportion of translations

• Amsili et al. (2012) investigate the pressure to use **additive markers**
  – Usually those are markers supposed to be obligatory
    • *Jo had fish, and Mo did # (too).*
  – There are “fringe cases” where the use of an additive appears optional:

  *Hartmann's joy was apparent in his beautifully cut hair, his expensive suit, his manicured hands, the faint aura of cologne that heralded his approach; in his mild and habitually smiling face, **too**, his expressive walk, in which the body, leaning slightly forward, seemed to indicate amiability*

• Does the pressure to use an additive varies cross-linguistically? Or is it a universal pragmatic constraint?
Preliminary results

• Comparison of the rate of non-translation of *also* and *only*

<table>
<thead>
<tr>
<th></th>
<th>Translation</th>
<th>Non-translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Also</em></td>
<td>157</td>
<td>65</td>
</tr>
<tr>
<td><em>Only</em></td>
<td>86</td>
<td>37</td>
</tr>
</tbody>
</table>

– The differences are not significant (Fisher’s test, $p = 0.9$)

• This (weakly) argues for a general account of the usage of discourse markers, consistent with some of the literature (Zeevat, 2014)

• Current work:
  – Distinguish between types of translations
  – Look at the CN/ No EN translations (annotation underway)
Argumentation algebra

• Argumentative operators are compositional (Winterstein, 2010)
  – Some combinations of markers are predicted to be more frequent than others:
    \( X \) *but only* \( X \)
    \( almost \) \( X \) *but did not* \( X \)

• Paired with a sentiment lexicon, the corpus can be used as a test bed for an argumentative algebra (e.g. Poria et al., 2014), under the assumption that utterances all argue for a similar goal
Further annotation

• Beyond the identification of the scope of a marker, the identification of full argumentative schemes is planned:
  – Argumentative premise
  – Argumentative conclusion
  – Type of argumentation (against/for, opposition, addition, parallelism etc.)

• This should help improve systems to automatically detect argumentation schemes
THANK YOU FOR YOUR ATTENTION
References