Identifying Entity Aspects in Microblog Posts

UNED NLP & IR Group
Madrid, Spain

damiano@lsi.uned.es, {edgar.meij, derijke}@uva.nl, {oghina,mbui,mbreuss}@science.uva.nl

ISLA, University of Amsterdam
Amsterdam, The Netherlands

35th International ACM SIGIR Conference on Research and Development in Information Retrieval
Portland, Oregon, USA. August 12–16, 2012.

Introduction

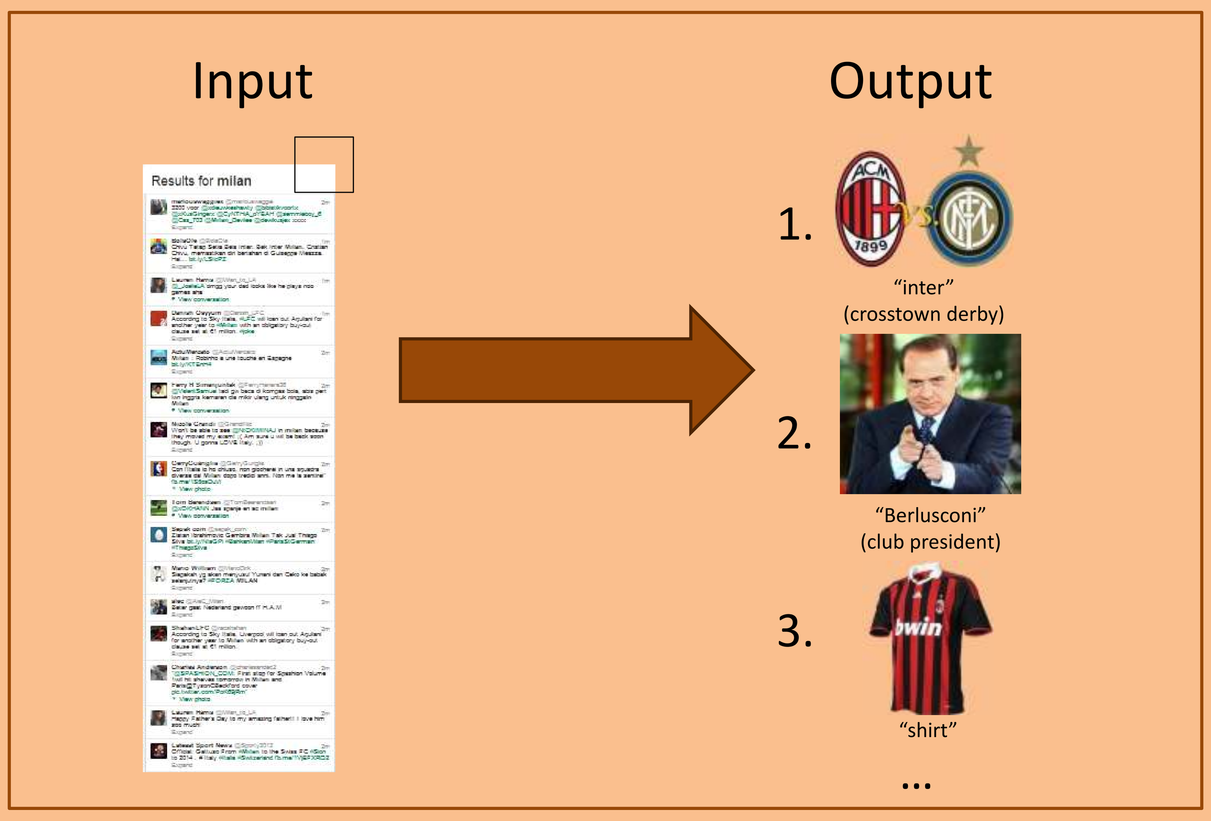
- Scenario: **Online reputation management**

  What do people say about an *entity*?
  entity = { company, organization, individual, product}

  Identification of aspects discussed on microblog posts

- **Aspects** ≈ products, services, competitors, key people related to the entity

Identifying Entity Aspects



Main idea

- Comparing a **pseudo-document *D*** built from entity-specific tweets with a **background corpus *C*.**
- Score of a term ***t* = *s(t, D, C)***

Four models tested:

- **TF.IDF**
- **LLR**: Log-Likelihood Ratio
- **PLM**: Parsimonious Language Models
- **OO**: Opinion-oriented. Opinion target extraction using topic-specific subjective lexicons

Method	Scoring function
TF.IDF	$s(t,D,C) = tf(t,D) \cdot \log \frac{N}{df(t)}$ N = number of pseudo-documents D_i in C $df(t) = \sum_i^N tf(t,D_i) > 0,\ D_i \in C$
LLR	$s(t,D,C) = 2 \cdot ((a \cdot \log(\frac{a}{E_1})) + (b \cdot \log(\frac{b}{E_2})))$ $E_1 = \frac{c \cdot (a+b)}{c+d} \quad E_2 = \frac{d \cdot (a+b)}{c+d}$ $a = tf(t,D) \quad b = cf(t)$ $c = \sum_i tf(t_i,D) \quad d = \sum_i cf(t_i)$
PLM	$s(t,D,C) = P(t\|D)$ (when model converges) E-step: $e_t = tf(t,D) \cdot \frac{\lambda \cdot P(t\|D)}{(1-\lambda)\cdot P(t\|C) + \lambda \cdot P(t\|D)}, \lambda = 0.1$ M-step: $P(t\|D) = \frac{e_t}{\sum_t e_t}$ initial $P(t\|D) = \frac{tf(t,D)}{\sum_i tf(t_i,D)} \quad P(t\|C) = \frac{cf(t)}{\sum_i cf(t_i)}$
OO	$s(t,D,C) = \chi^2(target(t,D), target(t,C))$ $\chi^2(o,e) = \frac{(o-e)^2}{e}$ $target(t,D)$ = freq. of potential target t in tweets D $target(t,C)$ = freq. of potential target t in background C

Dataset

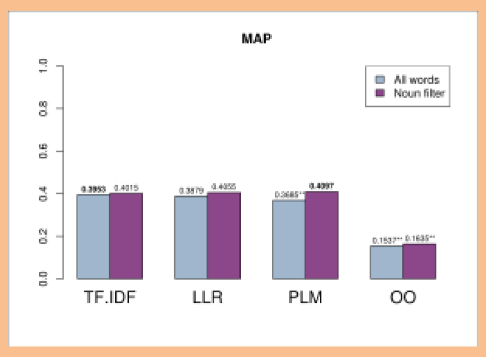
- 94 entities, 17,775 tweets, ≈177 tweets/entity
- Built upon WePS-3 ORM Task Dataset.
  Disambiguated company names (e.g. apple fruit vs. Apple Inc.)
- Pooling methodology.
  Three annotators with substantial agreement
  (Cohen/Fleiss' kappa > 0.6)
- Annotated 2455 terms, 1304 aspects (54.11%)
  Most of the true aspects are nouns (89.72%).

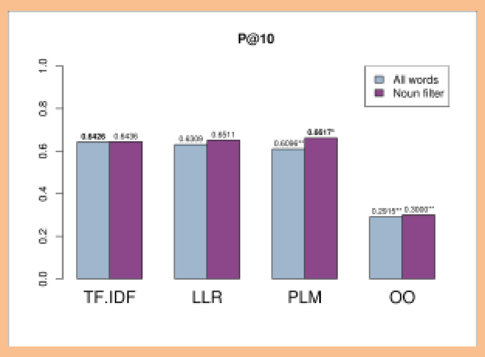
http://bit.ly/profilingTwitter

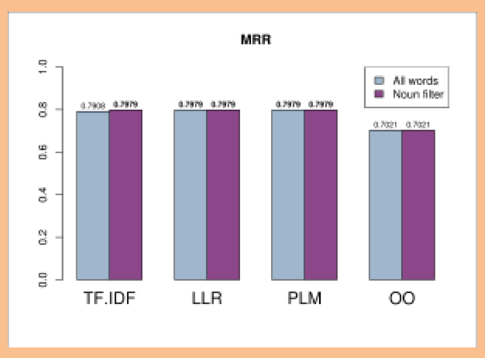
Entity	Aspects
Apple Inc.	ipad, iphone, prototype, store, gizmodo, employee
Sony	advertising, headphones, digital, pro, music, xperia, dsc, x10, bravia, camera, vegas, battery, ericsson, playstation
Starbucks	coffee, latte, tea, frappuccino, barista, drink, mocha

Experiments and Results

- **All Words:**
  - **TF.IDF** significantly **outperforms PLM** and **OO** in **precision**.
  - The results for **OO are much lower** than for the other methods.
- **Noun filter:**
  Applying a **part-of-speech filter** and only **consider terms tagged as nouns**
  - For all methods, MAP and precision values are slightly higher than in the all words condition: considering only nouns helps to identify aspects.
  - PLM best method using Noun filter.
- **Observations** (after manually inspecting the results):
  - Results for **TF.IDF, LLR** and **PLM** are **very similar**.
  - **OO** tends to return more **subjective terms as aspects** (syntactic parsing errors)
    Examples: haha, pls, xd, safety, win
  - **OO** has more **difficulty to filter out generic terms**.
    Examples: new, use, today, come







Student's t-test statistic significance w.r.t to TF.IDF *All words* baseline with α=0.05 (*) and α=0.01 (**).

Conclusions

- Simple statistical methods such as **TF.IDF** are a **strong baseline** for the task of **identifying entity aspects**, significantly **outperforming opinion-oriented methods**.
- **Only considering** terms tagged as **nouns improves** the **results** for **all the methods analyzed**.

Acknowledgments: This research was supported by the European Union's CIP ICT-PSP under grant agreement nr 250430, the European Community's FP7 Programme under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe), the Netherlands Organisation for Scientific Research (NOW) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, under COMMIT project Infiniti, the Spanish Ministry of Education (FPU grant nr AP2009-0507), the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02), the Regional Government of Madrid and the ESF under MA2VICMR (S2009/TIC-1542), the ESF Research Network Program ELIAS and the ACM Special Interest Group on Information Retrieval (SIGIR Student Travel Grant).