Combining retrieval, learning and NLP to estimate polarity in social media

Doctoral Meeting iniciative (CITIUS)

Jose M. Chenlo

josemanuel.gonzalez@usc.es
Centro de Investigación en Tecnoloxías da Información
CITIUS, University of Santiago de Compostela, Spain
www.gsi.dec.usc.es/ir

Advisor: David E. Losada

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

6 Future Work

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

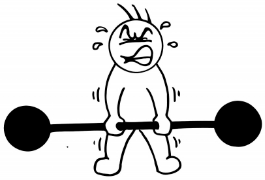
6 Future Work

Opinion-rich resources in Web

- Growing **availability** & **popularity**: online review sites, discussion forums, personal blogs, peer-to-peer networks, social networks, ...
- Opinions could be very ***valuable***: products/services, politics, **Reputation Management** ...
- ...***but*** most comercial search engines are not exploiting the opinionated nature of these sources of information
- OM & SA technology: potentially **wide industrial impact**

However,

- OM & SA technology still **not ready for prime time!**
- **Modest** levels of effectiveness



Topic retrieval vs Opinion retrieval

Topic retrieval: *easier* task

- Estimating topicality is somehow easier
- Keyword-based approaches work reasonably well
- Effective retrieval algorithms
- Massive success
  - Google
  - Yahoo!
  - Bing

Opinion retrieval: *harder* task

- Search for on-topic opinions: difficult passage-level task
- Locate key sentiments is challenging
- Deal with irony, sarcasm, etc.
- Context and Language dependent!
- Keyword-based approaches fail

Objective vs Subjective

Skype 2.0 eats its young

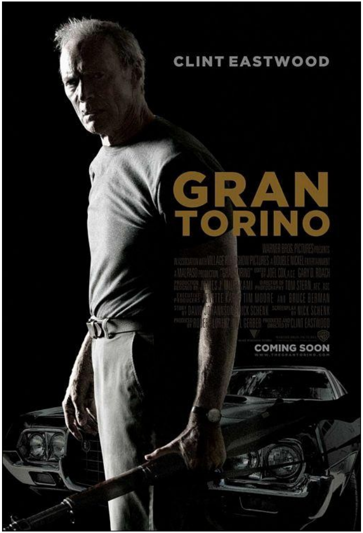
The elaborate press release and WSJ review while impressive don't help mask the fact that, Skype is short on new ground breaking ideas. Personalization via avatars and ring-tones · · · big new idea? Not really. Phil Wolff over on Skype Journal puts it nicely when he writes, "If you've been using Skype, the Beta version of Skype 2.0 for Windows won't give you a new Wow! experience."· · ·

Skype Launches Skype 2.0 Features Skype Video

Skype released the beta version of Skype 2.0, the newest version of its software that allows anyone with an Internet connection to make free Internet calls. The software is designed for greater ease of use, integrated video calling, and · · ·

Sentiment classification I

Gran Torino also includes a few easy outs built into the story ... And even without those easy outs, the storytelling's fairly obvious ... Gran Torino is a curdled mess, politically ... ***but considering that Gran Torino's heading towards the sunset of Eastwood's acting career, that's a good enough reason to watch it go by.***



Sentiment classification II

I hate the Spice Girls. . . . [3 things the author hates about them]. . . Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But. . . Okay, I'm really ashamed of it, but ***I enjoyed it***. I mean, I admit it's a really awful movie, . . . [they] act wacky as hell . . . the ninth floor of hell . . . a cheap [beep] movie . . . The plot is such a mess that it's terrible. But ***I loved it***.

State-of-the-art Summary: Main Issues

- In most cases, global (doc-level) methods
- Ignore the sequence of opinions
- Rough doc-level estimations
- Poor effectiveness in searching for pos & neg docs about a given topic

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

6 Future Work

Research Hypotheses

- There are **common patterns** in the way that people **express their sentiments**
- **Key sentiments** are located in **specific parts** of the text
- Search, learning and NLP can **guide the extraction** of key sentiments

Research Objectives

- ▷ **Location** of key sentiments
  - what are the **key parts of a document** for polarity estimation?
- ▷ Study of the **Sentiment Flow**
  - **How polarity evolves** throughout the document?
- ▷ Estimation of the **overall polarity**
  - How can we effectively and efficiently estimate the overall orientation of an opinionated text?

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

6 Future Work

Methodology: Model

- Search technology
  - State-of-the-art IR methods (BM25, LMs, ...) to extract on-topic passages
- Learning technologies
  - Supervised(e.g. sentiment classification) and unsupervised learning (e.g. clustering of common spans of text)
- NLP technologies
  - Extract linguistic features to improve search & learning
  - Understand & analyze the discourse structure

Methodology: Implementation

- ▷ Retrieval platform
  - Lucene, Lemur & Indri, Solr
- ▷ Learning platforms
  - Liblinear, Weka, ...
- ▷ NLP
  - SPADE ...
- ▷ ... and also our own developments & **extensions**

Methodology: Evaluation

- ▷ Large-scale benchmarks & standard measures
- ▷ Comparison against state-of-the-art approaches

- ▷ TREC Blog Track (**3,200k blogs**)
  - Multitopic
  - Labelled opinions at doc-level
  - Noisy collection (spam, off-topic, etc.)
- ▷ News collections (**4k sentences**)
  - More focused domain
  - Opinions labelled at sentence level
- ▷ Movie reviews (**1k reviews**)
  - Very focused
  - Small-scale
  - Labels at document level (# of stars)

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

6 Future Work

Publications

▷ Main publications

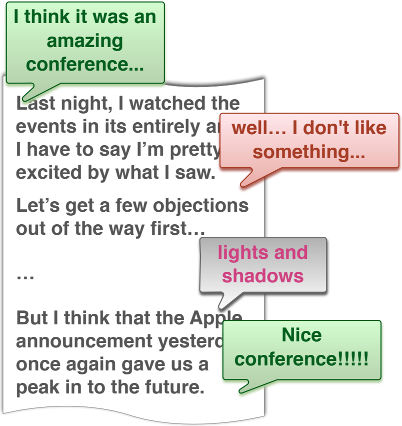
- **[Full Paper]** Jose M. Chenlo and David E. Losada. *Effective and Efficient Polarity Estimation in Blogs based on Sentence-Level Evidence*. In 20th ACM Conference on Information and Knowledge Management,CIKM 2011, Glasgow (Scotland), pages 365-374, 2011 ***(15 %acceptance rate)***. ***[CORE A]***
- **[Full Paper]** Jose M. Chenlo and David E. Losada. *Combining Document and Sentence Scores for Blog Topic Retrieval*. In 1st Spanish Conference on Information Retrieval , CERI 2010, Madrid (Spain), 2010 (46 % acceptance rate).
  - > Relevance model in a blog retrieval scenario.
- **[Full Paper]** J.M.Chenlo, J. Atserias, C. Rodriguez, R. Blanco. *FBM-Yahoo! at RepLab competition*. RepLab 2012 Lab, An evaluation campaign for Online Reputation Management Systems (within CLEF 2012), Rome (Italy), 2012.
  - > Publication in collaboration with Yahoo! Labs Barcelona
  - > Reputation Management over Twitter data.

Publications

- Main publications
  - **[Full Paper]** Jose M. Chenlo and David E. Losada. *Effective and Efficient Polarity Estimation in Blogs based on Sentence-Level Evidence*. In 20th ACM Conference on Information and Knowledge Management,CIKM 2011, Glasgow (Scotland), pages 365-374, 2011 ***(15 %acceptance rate)***. ***[CORE A]***
  - **[Full Paper]** Jose M. Chenlo and David E. Losada. *Combining Document and Sentence Scores for Blog Topic Retrieval*. In 1st Spanish Conference on Information Retrieval , CERI 2010, Madrid (Spain), 2010 (46 % acceptance rate).
    - > Relevance model in a blog retrieval scenario.
  - **[Full Paper]** J.M.Chenlo, J. Atserias, C. Rodriguez, R. Blanco. *FBM-Yahoo! at RepLab competition*. RepLab 2012 Lab, An evaluation campaign for Online Reputation Management Systems (within CLEF 2012), Rome (Italy), 2012.
    - > Publication in collaboration with Yahoo! Labs Barcelona
    - > Reputation Management over Twitter data.

Effective and Efficient Polarity Estimation in Blogs (I)

[challenge] what are the key parts of a document for polarity estimation?



Last night, I watched the events in its entirely a
I have to say I'm pretty
excited by what I saw.

Let's get a few objections out of the way first…

…

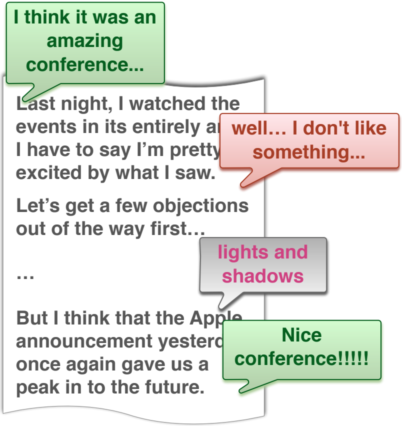
But I think that the Apple announcement yesterd
once again gave us a peak in to the future.

I think it was an amazing conference...
well… I don't like something...
lights and shadows
Nice conference!!!!!

Sentence-level analysis to obtain narrow bits of information with key evaluative statements

Effective and Efficient Polarity Estimation in Blogs (II)

[Main aim] Extract the **key parts** that represent the opinion of the writer about the query topic by using sentence-level evidence.



Last night, I watched the events in its entirely a I have to say I'm pretty excited by what I saw.

Let's get a few objections out of the way first...

...

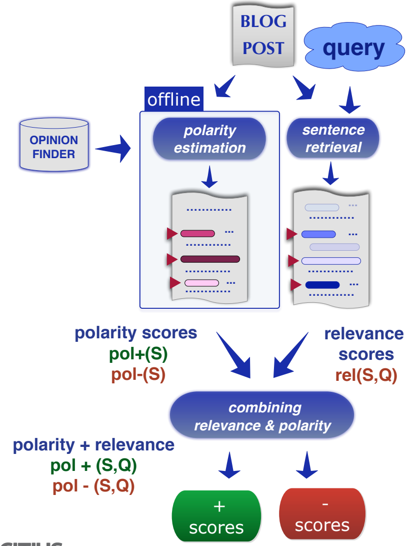
But I think that the Apple announcement yesterc once again gave us a peak in to the future.

I think it was an amazing conference...
well... I don't like something...
lights and shadows
Nice conference!!!!!

Sentence-level evidence

- ▷ Including:
  1. Relatedness of the sentence and the query topic
  2. Polarity of the sentence
     - > Opinion Finder
  3. **Relative location** of the subjective sentences

Effective and Efficient Polarity Estimation in Blogs (III)



PolMeanAll

Mean of polarity score of...
all on-topic polar (subjective) sentences

PolMeanBestN

Mean of polarity score of...
the *n* **most polar (subjective) on-topic sens**

PolMeanFirstN

Mean of polarity score of...
the *n* **first polar (subjective) sens**

PolMeanLastN

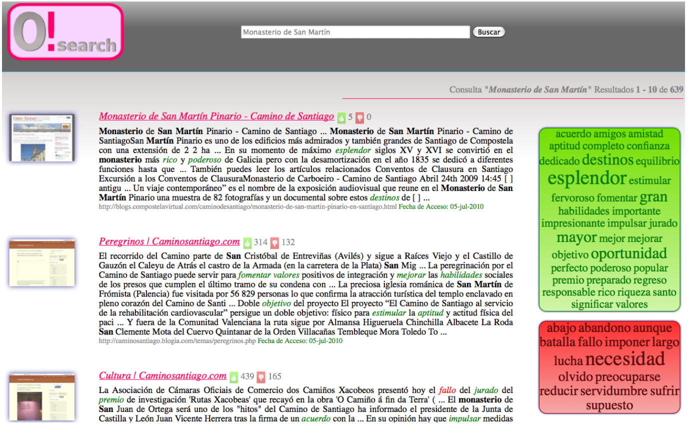
Mean of polarity score of...
the *n* **last polar (subjective) sens**

Effective and Efficient Polarity Estimation in Blogs (IV): Main findings

- Our sentence retrieval methods achieved state-of-the-art performance for polarity detection
- **first 4 polar sens** / **last 2 polar sens** are a good indicator of the overall polarity of a blog post
- Gains in efficiency
  - state-of-the-art approaches : need to process the whole doc
  - our location-aware methods : don't need to process the whole doc
  - "*Just take the first 4 polar (or the last 2 polar) sentences and classify the blog post accordingly"*

Other results: Software

- ▷ **O!Search**: A prototype application that retrieves opinions from blogs related to the "Way to Santiago".



Consulta "***Monasterio de San Martín***" Resultados **1** - **10** de **639**

Monasterio de San Martín Pinario - Camino de Santiago 👍 5 👎 0

Monasterio de **San Martín** Pinario - Camino de Santiago ... **Monasterio** de **San Martín** Pinario - Camino de SantiagoSan **Martín** Pinario es uno de los edificios más admirados y también grandes de Santiago de Compostela con una extensión de 2 2 ha ... En su momento de máximo *esplendor* siglos XV y XVI se convirtió en el **monasterio** más *rico* y *poderoso* de Galicia pero con la desamortización en el año 1835 se dedicó a diferentes funciones hasta que ... También puedes leer los artículos relacionados Conventos de Clausura en Santiago Excursión a los Conventos de ClausuraMonasterio de Carboeiro - Camino de Santiago Abril 24th 2009 14:45 [ ] antigu ... Un viaje contemporáneo" es el nombre de la exposición audiovisual que reune en el **Monasterio** de **San Martín** Pinario una muestra de 82 fotografías y un documental sobre estos *destinos* de [ ] ...
http://blogs.compostelavirtual.com/caminodesantiago/monasterio-de-san-martin-pinario-en-santiago.html Fecha de Acceso: 05-jul-2010

Peregrinos | Caminosantiago.com 👍 314 👎 132

El recorrido del Camino parte de **San** Cristóbal de Entreviñas (Avilés) y sigue a Raíces Viejo y el Castillo de Gauzón el Caleyu de Atrás el castro de la Armada (en la carretera de la Plata) **San** Mig ... La peregrinación por el Camino de Santiago puede servir para *fomentar valores* positivos de integración y *mejorar* las *habilidades* sociales de los presos que cumplen el último tramo de su condena con ... La preciosa iglesia románica de **San Martín** de Frómista (Palencia) fue visitada por 56 829 personas lo que confirma la atracción turística del templo enclavado en pleno corazón del Camino de Santi ... Doble *objetivo* del proyecto El proyecto "El Camino de Santiago al servicio de la rehabilitación cardiovascular" persigue un doble objetivo: físico para *estimular* la *aptitud* y actitud física del paci ... Y fuera de la Comunitad Valenciana la ruta sigue por Almansa Higueruela Chinchilla Albacete La Roda **San** Clemente Mota del Cuervo Quintanar de la Orden Villacañas Tembleque Mora Toledo To ...
http://caminosantiago.blogia.com/temas/peregrinos.php Fecha de Acceso: 05-jul-2010

Cultura | Caminosantiago.com 👍 439 👎 165

La Asociación de Cámaras Oficiais de Comercio dos Camiños Xacobeos presentó hoy el *fallo* del *jurado* del *premio* de investigación 'Rutas Xacobeas' que recayó en la obra 'O Camiño á fin da Terra' (... El **monasterio** de **San** Juan de Ortega será uno de los "hitos" del Camino de Santiago ha informado el presidente de la Junta de Castilla y León Juan Vicente Herrera tras la firma de un *acuerdo* con la ... En su opinión hay que *impulsar* medidas

acuerdo amigos amistad aptitud completo confianza dedicado destinos equilibrio esplendor estimular fervoroso fomentar gran habilidades importante impresionante impulsar jurado mayor mejor mejorar objetivo oportunidad perfecto poderoso popular premio preparado regreso responsable rico riqueza santo significar valores

abajo abandono aunque batalla fallo imponer largo lucha necesidad olvido preocuparse reducir servidumbre sufrir supuesto

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

6 Future Work

Main difficulties

- Web is noisy
  - Spam
  - Off-topic material
  - On-topic material without opinion
  - Opinions non related to the query
- Irony
- Context-dependent
- Language-dependent
- In general, performance issues
  - Big collections (3,200k pages from blogs crawled from the web)
  - Advanced NLP methods, such as RST required a lot of processing time.

Challenges

- ▷ Extracting key passages for sentiment prediction is good...
- ▷ ***BUT***, we also need a formal & expressive way to capture the sequence of sentiments (location, strength & context of each on-topic opinion)
- ▷ Understanding how polarity evolves throughout the document is essential for sentiment prediction

Ongoing research

> Although it was great to see Brad Pitt fall off a cliff, ***this movie was terrible***

- ▷ Polarity estimation using (sentence-level) discourse structure
- ▷ Rhetorical Structure Theory (RST):
  - Sentences split into nucleus + satellite
- ▷ Different rhetorical relations: attribution, background, cause, elaboration, ...
- ▷ Preliminary results found ***to be published*** in collaboration with the Erasmus University of Rotterdam (Alexander Hogenboom)

Outline

1 Context & Motivation

2 Research Hypotheses & Objectives

3 Methodology

4 Current Results

5 Challenges

6 Future Work

Future Work

- To study more **refined ways of representing the flow of sentiments**
  - Evolved discourse analysis
  - **Intra-sentence** analysis
  - **Inter-sentence** sentiment flow
- **Machine learning** to locate key sentiments in a context-dependent way
- Integrated **search for on-topic opinions**
- Explore the creation of **opinion-biased summaries**

Thank you!!

Opinions? :-)