An example of application: Sense-based user model

- SiteIF is a personal agent for a news web site that learns user's interests from the requested pages that are analyzed to generate or to update a model of the user.
- Exploiting this model, the system anticipates which documents in the web site could be interesting for the user.
- (Magnini, Strapparava 2004)
The news corpus

- A bilingual (English/Italian) web site of a news provider (Adnkronos)
- The average length of each news is about 256 words. => “One domain per discourse” assumption is mainly fulfilled

Parallel Texts

CULTURE: GIOTTO- PAID BY MONKS TO WRITE ANTI-FRANCISCAN POETRY
Rome; 10 Jan. - (Adnkronos) - Giotto was ‘paid’ to attack a faction of the Franciscans, the Spiritual ones, who opposed church decoration in honour of Poverello di Assisi. This has been revealed in the research of an Italian scholar who is a professor at Yale University, Stefano Ugo Baldassarri, who thinks he has solved the mystery of the only known poetry by the famous Tuscan painter: the Giotto verses have in fact always provoked wonder because they seem to be a criticism of the ideals of St. Francis and all the more so since their author was also the man who painted the famous frescoes of the Basilica at Assisi….

CULTURA: GIOTTO- PAGATO DA FRATI PER SCRIVERE POESIA ANTI-FRANCESCAN
Roma; 10 gen. - (Adnkronos) - Giotto fu ‘pagato’ per attaccare una fazione dei Francescani, quella degli Spirituali, che si opponevano alla decorazione delle chiese in onore del Poverello di Assisi. Lo rivela una ricerca di uno studioso italiano docente alla Yale University, Stefano Ugo Baldassarri, che ritiene di aver svelato il mistero dell'unica poesia conosciuta del celebre pittore toscano: i versi giotteschi, infatti, avevano sempre destato meraviglia perché apparivano come una critica agli ideali di San Francesco, tanto più mossa proprio dall'autore dei celebri affreschi della Basilica di Assisi….
A sense-based user model

- Limitation of many actual system:
  - the technique used to represent a user’s profile is often based on simple lists of keywords (=> polysemy problems ...)
  - The idea => Build the user model as a network whose nodes are synsets from MultiWordNet
  - We use Domain Driven Disambiguation => good performance on “domain” words

---

Document Representation (using a WSD technique)

<table>
<thead>
<tr>
<th>File: t0003</th>
<th>Tagged File: t0003.tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Polisemy: 3.892857</td>
<td>Total tokens: 102</td>
</tr>
<tr>
<td>Best Scored SemField (BSSF): literature</td>
<td>Total tokens in the BSSF: 16</td>
</tr>
<tr>
<td>Proper Noun List: Giotto, Assisi,...</td>
<td>List of synsets belonging to BSSF: poetry#1, author#1, ...</td>
</tr>
<tr>
<td>Non ambiguous synsets list: Franciscan#1, controversy#1,...</td>
<td></td>
</tr>
</tbody>
</table>

---

CULTURE: GIOTTO - PAID BY MONKS TO WRITE ANTI-FRANCISCAN POETRY

Rome, 10 Jan. (Adnkronos) - Giotto was ‘paid’ to attack a faction of the Franciscans, the Spiritual ones, who opposed church decoration in honour of Poverello di Assisi. This has been revealed in the research of an Italian scholar who is a professor at Yale University, Stefano Ugo Baldassarri, who thinks he has solved the mystery of the only known poetry by the famous Tuscan painter: the Giotto verses have - in fact always provoked wonder because they seem to be a criticism of the ideals of St. Francis and all the more so since their author was also the man who painted the famous frescoes of the Basilica at Assisi.
Consequences for User Model

- Not just the “same” words as other visited documents, but the same concepts
- The user interest model is independent from the language of the documents passed over by the user
- It would also be possible to introduce a notion of user model coherence => how much the user model talks “about the same things”
- Dynamically inferring the interest areas of the resulting user model
WSD and applications

- We don’t have a “perfect” WSD algorithm at the moment
- Different WSD algorithms and techniques have different characteristics
- Ex. DDD is focused on domains - it disambiguates well the “domain” words
- Select which algorithm is better for a specific application or develop a new one

Applying WSD - Future directions

- Some applications where “traditional” WSD is only one component and the whole system requires a deep understanding at sense-level.
- Educational systems
- Help to learn a second-language (how to evaluate this?)
- Creative NLP systems (Stock & Strapparava - IJCAI-2003)