Speech Act Based Classification of Email Messages in Croatian Language

Tin Franović, Jan Šnajder
{tin.franovic,jan.snajder}@fer.hr

Eighth Language Technologies Conference (LTC IS-2012)
Ljubljana, October 8th, 2012
Background & motivation

- Increase in popularity of email as means of communication
- Recent surveys – up to 2 hours a day spent on emails
- Automated email classification can reduce the amount of time users spend reading and sorting emails

Speech acts (Searle, 1965)

Speech acts are illocutionary acts that attempt to convey meaning from the speaker (or writer) to the listener (or reader)

- Speech acts are effective way of summarizing the intended purpose of the message
Background & motivation

- Increase in popularity of email as means of communication
- Recent surveys – up to 2 hours a day spent on emails
- Automated **email classification** can reduce the amount of time users spend reading and sorting emails

Speech acts (Searle, 1965)

Speech acts are illocutionary acts that attempt to convey meaning from the speaker (or writer) to the listener (or reader)

- Speech acts are effective way of **summarizing the intended purpose** of the message
Our goal

Develop and evaluate **speech act classification** of email messageg in Croatian language using **supervised machine learning**

- Task framed as a **multilabel text classification problem**
- Thorough evaluation using six machine learning algorithms
- Evaluated using message-level, paragraph-level, and sentence-level features
Our goal

Develop and evaluate speech act classification of email messages in Croatian language using supervised machine learning

- Task framed as a multilabel text classification problem
- Thorough evaluation using six machine learning algorithms
- Evaluated using message-level, paragraph-level, and sentence-level features
Coming up next... 

1. Message classification
   - Dataset
   - Message preprocessing
   - Training classifiers

2. Evaluation

3. Conclusion and future work
Several publicly available email datasets, however none in Croatian

We compiled a dataset using 1337 messages from five sources

Annotated using 13 different speech acts [Searle, 1965]

- Assertives (AMEND, PREDICT, CONCLUDE);
- Directives (REQUEST, REMIND, SUGGEST);
- Expressives (APOLOGIZE, GREET, THANK);
- Commisives (COMMIT, REFUSE, WARN);
- Declarations (DELIVER).
Two annotators, 15% of dataset double-annotated

<table>
<thead>
<tr>
<th>Speech act</th>
<th>$\kappa$</th>
<th>Speech act</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMEND</td>
<td>0.714</td>
<td>REFUSE</td>
<td>0.000</td>
</tr>
<tr>
<td>APOLOGIZE</td>
<td>0.856</td>
<td>REMIND</td>
<td>0.747</td>
</tr>
<tr>
<td>COMMIT</td>
<td>0.851</td>
<td>REQUEST</td>
<td>0.589</td>
</tr>
<tr>
<td>CONCLUDE</td>
<td>0.005</td>
<td>SUGGEST</td>
<td>0.544</td>
</tr>
<tr>
<td>DELIVER</td>
<td>0.792</td>
<td>THANK</td>
<td>0.949</td>
</tr>
<tr>
<td>GREET</td>
<td>0.779</td>
<td>WARN</td>
<td>0.174</td>
</tr>
<tr>
<td>PREDICT</td>
<td>0.267</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dataset annotation

- Infrequent and low-IAA speech acts removed:
  - APOLOGIZE, CONCLUDE, GREET, PREDICT, REFUSE, THANK, WARN
- Speech acts used:
  - DELIVER, AMEND, COMMIT, REMIND, SUGGEST, REQUEST

Pozdrav Marko!
Molim te da mi što prije pošalješ novu verziju sa izmijenjenim podacima. Treba dodati vrijeme i datum.
Hvala ti!
Message preprocessing

- Reduce the dimensionality and morphological variation
- Stemming
  - Suffix of each word after last vowel removed
  - Number of terms reduced from 15,100 to 11,856
- Stop-word removal
  - Filtered out words with little semantic information
  - List of 2,024 Croatian stop-words
Message preprocessing (2)

- Separate training set created for each speech act using annotated data
- Text segments extracted at corresponding discourse levels
  - Sentence and paragraph levels – segments that enclose start and end point of annotation
  - Message level – complete message
- Negative examples sampled from the set of segments not annotated with the corresponding speech act
Training classifiers

- Rapid Miner implementation
- **Six different models:**
  - SVMs (Support Vector Machines), naive Bayes (NB), k-NN (\(k\)-Nearest Neighbors), Decision Stump (DS), AdaBoost (with Decision Stump as the weaker learner), and RDR (Ripple Down Rule)
- **Three term weighting schemes:**
  - TF (Term Frequency) and TF-IDF (Term Frequency – Inverted Document Frequency) - all models except RDR
  - Binary weights - only RDR
- Separate classifier trained for every speech act, term weighting scheme, and discourse level (198 models)
- Re-trained using stop-word removal
Training classifiers (2)

- Parameter optimization
  - Grid-search
  - 10-fold cross-validation for every parameter combination
  - Optimal parameter chosen based on averaged F1 score
- Optimal model re-trained using whole training set and tested on held-out set
- 70% for training/validation, 30% held-out test set
F1 performance for best feature/discourse level combinations:

<table>
<thead>
<tr>
<th></th>
<th>NB</th>
<th>k-NN</th>
<th>SVM</th>
<th>DS</th>
<th>AB</th>
<th>RDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliver</td>
<td>69.70</td>
<td>83.72</td>
<td>88.16</td>
<td>85.71</td>
<td>87.50</td>
<td><strong>88.51</strong></td>
</tr>
<tr>
<td>Amend</td>
<td><strong>79.31</strong></td>
<td>71.43</td>
<td>77.97</td>
<td>72.29</td>
<td>74.63</td>
<td>77.27</td>
</tr>
<tr>
<td>Commit</td>
<td>62.45</td>
<td>67.44</td>
<td>78.61</td>
<td>79.37</td>
<td>81.97</td>
<td><strong>83.75</strong></td>
</tr>
<tr>
<td>Remind</td>
<td>60.87</td>
<td>63.64</td>
<td>75.00</td>
<td>76.92</td>
<td><strong>94.74</strong></td>
<td>76.92</td>
</tr>
<tr>
<td>Suggest</td>
<td>67.06</td>
<td>70.27</td>
<td><strong>76.84</strong></td>
<td>76.27</td>
<td>75.12</td>
<td>71.50</td>
</tr>
<tr>
<td>Request</td>
<td>69.69</td>
<td>75.44</td>
<td><strong>78.76</strong></td>
<td>70.57</td>
<td>75.23</td>
<td>74.46</td>
</tr>
</tbody>
</table>

UNIZG FER TakeLab | October 8th, 2012
F1 performance for best classifier/feature combinations:

<table>
<thead>
<tr>
<th></th>
<th>Message</th>
<th>Paragraph</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliver</td>
<td>86.59</td>
<td>83.64</td>
<td>88.51</td>
</tr>
<tr>
<td>Amend</td>
<td>79.31</td>
<td>77.27</td>
<td>72.38</td>
</tr>
<tr>
<td>Commit</td>
<td>83.75</td>
<td>81.97</td>
<td>78.93</td>
</tr>
<tr>
<td>Remind</td>
<td>94.74</td>
<td>76.92</td>
<td>69.57</td>
</tr>
<tr>
<td>Suggest</td>
<td>71.88</td>
<td>76.84</td>
<td>69.74</td>
</tr>
<tr>
<td>Request</td>
<td>70.09</td>
<td>78.76</td>
<td>72.19</td>
</tr>
<tr>
<td>Overall</td>
<td>94.74</td>
<td>83.64</td>
<td>78.93</td>
</tr>
</tbody>
</table>
Feature types

- F1 performance for best classifier/discourse level combinations:

<table>
<thead>
<tr>
<th>Feature</th>
<th>With stop-words</th>
<th>Without stop-words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary</td>
<td>TF</td>
</tr>
<tr>
<td>Deliver</td>
<td>88.51</td>
<td>87.50</td>
</tr>
<tr>
<td>Amend</td>
<td>70.07</td>
<td>77.19</td>
</tr>
<tr>
<td>Commit</td>
<td>83.75</td>
<td>79.37</td>
</tr>
<tr>
<td>Remind</td>
<td>76.92</td>
<td>76.92</td>
</tr>
<tr>
<td>Suggest</td>
<td>71.50</td>
<td>76.84</td>
</tr>
<tr>
<td>Request</td>
<td>61.90</td>
<td>78.76</td>
</tr>
</tbody>
</table>
## Overall performance

- F1 performance with optimal feature sets for each classifier, averaged over speech acts:

<table>
<thead>
<tr>
<th></th>
<th>Message</th>
<th>Paragraph</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>79.31</td>
<td>69.70</td>
<td>72.38</td>
</tr>
<tr>
<td>k-NN</td>
<td>72.73</td>
<td>75.44</td>
<td>83.72</td>
</tr>
<tr>
<td>SVM</td>
<td>83.87</td>
<td>81.55</td>
<td>88.16</td>
</tr>
<tr>
<td>DS</td>
<td>78.65</td>
<td>79.37</td>
<td>85.71</td>
</tr>
<tr>
<td>AB</td>
<td>94.74</td>
<td>83.54</td>
<td>87.50</td>
</tr>
<tr>
<td>RDR</td>
<td>86.59</td>
<td>83.64</td>
<td>88.51</td>
</tr>
</tbody>
</table>
Conclusion

- Addressed multilabel speech act classification for Croatian
- Thorough evaluation using six machine learning algorithms and three feature types
- Discourse level and feature type do not influence significantly classification performance
- Certain speech acts more accurately classified on particular levels
- Obtained F1 scores notably higher than reported in previous work [Cohen, 2004; Carvalho, 2006]
Future work

- Explore relationship between discourse level and speech acts
- Employ information extraction methods to augment speech acts
- Impact of speech acts on importance-based classification
Thank you for your attention

Let’s keep in touch. . .

www.takelab.hr

info@takelab.hr