# Introducing the Notion of 'Contrast' Features for LanguageTechnology

Marina Santini, Benjamin Danielsson, **Arne Jönsson**

RISE Research Institutes of Sweden

Div. ICT-RISE SICS East

# Outline

- Genre and domain
- Contrast features
- Experiments
- Conclusion

# Genre or Domain? Sorting out Text Varieties

- **Domain** is a subject field: e.g. ”Fashion”, ”Leisure”, ”Business”, ”Sport”, ”Medicine” or ”Education”. In text classification, domains are normally represented by topical features, such as content words and specialized terms.
- **Genre** refers to conventionalized textual patterns, e.g. ”academic papers”, ”tweets”, ”letters” and ”interviews”. In text classification, genres are often represented by features such as POS tags, character n-grams, POS n-grams, syntactic tags and function words.

# How can we automatically separate genre from domain?

- Is it possible to decide automatically whether a text category is a genre or a domain?
- We explore whether there exist ’contrast’ features that help recognize if a text category is a genre or a domain.

# ’Contrast’ Features

- ’Contrast’ features are features that consistently perform well (or badly) only on either genre or domain.
- We experiment with the text categories included in the Swedish National Corpus (SUC).
- We build text classification models based on different feature sets
- Which one(s) of these feature sets are the most reliable ‘contrast’ features?

# Experiments

- Three sets of experiments based on features:
  1. text complexity- and grammatical features
  2. BoW features
  3. function words and word embeddings.
- Supervised machine learning (weka)
  - Support Vector Machines (SMO)
  - Multilayer Perceptron (DI4jMlp)
- Weighted Averaged F-Measure
  - 10-folds cross validation

# SUC Text Categories

The SUC is a collection of Swedish texts (amounting to about one million words) and represents the Swedish language of the 1990's

9 text categoires:

a. reportage genre
b. editorial genre
c. review genre
d. hobby domain
e. popular lore domain
f. bio essay genre
g. miscellaneous mixed
h. scientific writing genre
i. imaginative prose genre

# Text complexity features

- Shallow features
  - Extracted after tokenization by simply counting words and characters.
- Lexical features
  - Based on categorical word frequencies extracted after lemmatization and calculated using the basic Swedish vocabulary SweVoc
- Morpho-syntactic features
  - Based on a morphology  analysis of text
- Syntactic features
  - Features estimable after syntactic parsing of the text
- Text quality metrics
  - Metrics used to measure readability for Swedish

# Principal component analysis (PCA)

- Each component comprises parameters with varying weight

| ent | Parametral | Komponenten | Förklaring |
|---|---|---|---|
| 1 | avgNominalPostmodifiers | ,855 | The average number of nominal postmodifiers per sentence |
| | avgNominalPremodifiers | ,537 | The average number of nominal premodifiers per sentence |
| | avgPrepComp | ,860 | The average number of prepositional complements per sentence in the document |
| | avgSentenceDepth | ,739 | The average sentence depth |
| | avgSentenceLength | ,944 | The average sentence length |
| | avgWordsPerClause | ,867 | The average number of words per clause in the document. |
| | dep_ET | ,439 | Other nominal post-modifier |
| | dep_IP | -,751 | Period |
| | dep_SS | -,435 | Other subject |
| | lixValue | ,734 | Läsbarhetsindex, readability index. |
| | meanDepDistanceDependent | ,778 | The mean dependency distance in the document on a per dependent basis. |
| | meanDepDistanceSentence | ,787 | The mean dependency distance in the document on a per sentence basis. |
| | nrValue | -,448 | Nominal ratio |
| | pos_MAD | -,752 | MAD Major delimiter (.?!) |
| 2 | dep_SS | ,432 | Other subject |
| | dep_UA | ,920 | Subordinate clause minus subordinating conjunction |
| | pos_SN | ,882 | SN Subjunction |
| | pos_VB | ,449 | VB Verb |
| | verbArity2 | ,404 | The ratios of verbs with an arity of 0-7, that is, the ratio of verbs with an arity of 0 as one f |
| 3 | dep_IK | ,799 | Comma |
| | dep_IT | ,608 | Dash |

# Set 1
# Experiments with text complexity features and grammatical features

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 9 SUC varieties (a reportage_genre, b editorial_genre,c review_genre, e hobby_domain, f popular_lore_domain, g bio_essay_genre, h miscellaneous_mixed, j scientifc _writing_genre, k imaginative_pros _genre) 1400 instances | 115 complexity features | **0.596** | 0,582 |
| | 65 components | 0.567 | 0.572 |
| | 27 POS tags | 0.507 | 0.526 |
| | 62 dependency tags | 0.541 | 0.531 |

# Set 1 Confusion matrix
# 9 text categories SMO: F 0,596

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 209   3   6  24   2   0  15   1   9 |   a = a reportage genre
   8  36   5  10   3   0   3   4   1 |   b = b editorial genre
  13   4 100   4   0   0   0   4   2 |   c = c review genre
  65   4   7  23   5   1   7   3   9 |   d = e hobby domain
  10   5   2   9  17   0   6   8   5 |   e = f popular lore domain
   2   3   3   2   0   3   0   8   6 |   f = g bio essay genre
  42   8   1   9   6   1  70   7   1 |   g = h miscellaneous mixed
   4   3   3   3   3   0   5  63   2 |   h = j scientific writing genre
   5   0   0   2   0   1   0   0 122 |   i = k imaginative prose genre
```

# Set 1: ‘Proper genres’

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 5 SUC genres (a reportage_genre, b editorial_genre,c review_genre, j scientific_writing_genre, k imaginative_pros _genre) 682 instances | 115 complexity features | **0.831** | 0.531 |
| | 65 components | 0.829 | 0.811 |
| | 27 POS tags | 0.786 | 0.773 |
| | 62 dependency tags | 0.782 | 0.771 |

# Set 1. 2 genres and 2 domains

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 4 SUC varieties (2 domains and 2 genres; e hobby_domain, f popular_lore_domain, j scientific _writing_genre, k imaginative_pros _genre) 402 instances | 115 complexity features | **0.785** | 0.766 |
| | 65 components | 0.722 | 0.704 |
| | 27 POS tags | 0.743 | 0.740 |
| | 62 dependency tags | 0.715 | 0.711 |

# Set 1. 2 genres vs 2 domains

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 2 SUC genres (j scientific _writing_genre, k imaginative_pros _genre) 216 instances | 115 complexity features | 0.981 | 0.981 |
| | 65 components | 0.972 | 0.949 |
| | 27 POS tags | **0.986** | 0.981 |
| | 62 dependency tags | 0.981 | 0.968 |
| | | | |
| **SUC Text Categories** | **Features** | **SMO** | **DI4jMlp** |
| 2 SUC domains (e hobby_domain, f popular_lore_domain) 186 instances | 115 complexity features | 0.720 | **0.749** |
| | 65 components | 0.692 | 0.674 |
| | 27 POS tags | 0.674 | 0.706 |
| | 62 dependency tags | 0.707 | 0.722 |

# Set 2
# Experiment with bag-of-words features

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 9 SUC varieties (a reportage_genre, b editorial_genre,c review_genre, e hobby_domain, f popular_lore_domain, g bio_essay_genre, h miscellaneous_mixed, j scientifc _writing_genre, k imaginative_pros _genre) 1400 instances | Including stopwords | **0.767** | 0.640 |
| | Without stopwords | 0.741 | 0.614 |

# Set 2. Confusion matrix BoWs

## 9 text categories SMO: F 0,767 (with stopwords)

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 231   5   5  21   1   0   3   1   2 |   a = a_reportage_genre
  14  40   2  10   0   0   4   0   0 |   b = b_editorial_genre
  17   0 108   0   0   0   0   1   1 |   c = c_review_genre
  31   7   5  68   5   0   4   2   2 |   d = e_hobby_domain
   4   2   0   6  39   0   5   6   0 |   e = f_popular_lore_domain
   0   2   0   0   2   8   0   6   9 |   f = g_bio_essay_genre
  22   1   0  11   1   0 108   2   0 |   g = h_miscellaneous_mixed
   0   0   1   3   2   0   3  77   0 |   h = j_scientific_writing_genre
   1   0   3   1   0   1   0   0 124 |   i = k_imaginative_prose_genre
```

# Set 1: ‘Proper genres’

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 5 SUC genres (a reportage_genre, b editorial_genre,c review_genre, j scientific_writing_genre, k imaginative_pros _genre) 682 instances | Including stopwords | **0.903** | 0.854 |
| | Without stopwords | 0.863 | 0.824 |

# Set 1. 2 genres and 2 domains

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 4 SUC varieties (2 domains and 2 genres; e hobby_domain, f popular_lore_domain, j scientific _writing_genre, k imaginative_pros _genre) 402 instances | Including stopwords | **0.905** | 0.828 |
| | Without stopwords | 0.880 | 0.792 |

# Set 1. 2 genres vs 2 domains

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 2 SUC genres (j scientific _writing_genre, k imaginative_pros _genre) 216 instances | Including stopwords | **0.991** | **0.991** |
| | Without stopwords | **0.991** | **0.991** |

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 2 SUC domaina (e hobby_domain, f popular_lore_domain) 186 instances | Including stopwords | **0.925** | 0.858 |
| | Without stopwords | 0.892 | 0.842 |

# Set 3. Function Words (15 POS tags) vs Word2Vec Word Embeddings

| SUC Text Categories | Features | SMO | DI4jMlp |
|---|---|---|---|
| 9 SUC varieties (a reportage_genre, b editorial_genre,c review_genre, e hobby_domain, f popular_lore_domain, g bio_essay_genre, h miscellaneous_mixed, j scientifc _writing_genre, k imaginative_pros _genre) 1400 instances | Function words | 0.371 | **0.448** |
| | Word Embeddings | n/a | 0.340 |

# Summary

- *Text complexity features and grammatical features* **do have  the contrastive power to disentangle genres and domains**. They are more representative of genres than domains and mixed classes, since they perform consistently better on genre classes.
- BoW features perform equally well on genres and on domains. They **do not have contrastive power**.
- Function words and word embeddings have a weak overall performance on the SUC.

# Conclusion

Text complexity features and grammatical features are more suitable as ’contrast’ features than BoW features.

# Future Work

- Exploration of additional ’contrast’ features
- Further exploration of their effectiveness on other corpora.

# Acknowledgements

This research was supported by E-care@home, a “SIDUS – Strong Distributed Research Environment” project, funded by the Swedish Knowledge Foundation.