Word-based dialect identification with georeferenced rules

SCHERRER, Yves

Available at: http://archive-ouverte.unige.ch/unige:22809

Disclaimer: layout of this document may differ from the published version.
Word-based Dialect Identification
With Georeferenced Rules

Yves Scherrer
Université de Genève

26.10.2010
Introduction

Some Swiss German examples:

http://als.wikipedia.org/wiki/Alemannischer_Beispielsatz

\[
\begin{align*}
Bisch uff em Märt gsi go yykaufe? & \rightarrow BA \\
Bisch uf e Märit ga kömerle? & \rightarrow BE \\
Büschn z’Määret gsi ga iichuke? & \rightarrow FR \\
Bisch uf dä Märt go poschtä? & \rightarrow OS \\
Bisch dü uf um Markt gsii ga ichöifu? & \rightarrow WS \\
Bisch uf em Mërt gsy go poschte? & \rightarrow ZH \\
Warst du auf dem Markt einkaufen? & \rightarrow Standard German
\end{align*}
\]

Characteristics:

- No clear-cut dialect boundaries.
- We use written dialect data.
- No standardized spelling.
Introduction

Some Swiss German examples:

http://als.wikipedia.org/wiki/Alemannischer_Beispielsatz

Bisch uff em Märt gsi go yykaufe? → BA
Bisch uf e Märit ga kömerle? → BE
Büschu z’Määret gsi ga iichufe? → FR
Bisch uf dä Märt go poschtä? → OS
Bisch dü uf um Markt gsii ga ichöifu? → WS
Bisch uf em Mèrt gsy go poschte? → ZH
Warst du auf dem Markt einkaufen? → Standard German

Characteristics:

- No clear-cut dialect boundaries.
- We use written dialect data.
- No standardized spelling.
Approaches

1. **N-gram model, trained**
   - A simple model successfully used for language identification: Learn character n-gram distributions in a supervised approach
     - Dialect ID $\approx$ Language ID
   - How does this model handle the specificities of dialects?
     - Data sparseness issues with training corpora from different dialects
     - Dialects are closely related to each other, and thus more difficult to distinguish

2. **Word-based model, using external knowledge**
   - Specific words carry more precise localization information than the n-gram distribution does
   - Use external resources to overcome the training data bottleneck
     - Linguistic atlases contain detailed information about dialect differences and allow us to model continuous language change
     - Dialect ID $\neq$ Language ID
Approaches

1. **N-gram model, trained**
   - A simple model successfully used for language identification: Learn character n-gram distributions in a supervised approach
     - Dialect ID $\approx$ Language ID
   - How does this model handle the specificities of dialects?
     - Data sparseness issues with training corpora from different dialects
     - Dialects are closely related to each other, and thus more difficult to distinguish

2. **Word-based model, using external knowledge**
   - Specific words carry more precise localization information than the n-gram distribution does
   - Use external resources to overcome the training data bottleneck
     - Linguistic atlases contain detailed information about dialect differences and allow us to model continuous language change
     - Dialect ID $\neq$ Language ID
Approaches

1. **N-gram model, trained**
   - A simple model successfully used for language identification:
     - Learn character n-gram distributions in a supervised approach
     - **Dialect ID \( \approx \) Language ID**
   - How does this model handle the specificities of dialects?
     - Data sparseness issues with training corpora from different dialects
     - Dialects are closely related to each other, and thus more difficult to distinguish

2. **Word-based model, using external knowledge**
   - Specific words carry more precise localization information than the n-gram distribution does
   - Use external resources to overcome the training data bottleneck
     - Linguistic atlases contain detailed information about dialect differences and allow us to model continuous language change
     - **Dialect ID \( \neq \) Language ID**
1. Introduction

2. Data

3. N-Gram model

4. Word-based model
   - The idea
   - Creating georeferenced rules
   - The identification process
   - Extensions

5. Conclusion
Data

Need texts that are annotated with their dialect:

- **Swiss German** Wikipedia

Other **Web** texts whose dialect could be inferred:
Websites and blogs of local sports clubs, music bands etc.
Dialects

Yves Scherrer: Word-based dialect identification with georeferenced rules
Data

Number of sentences per dialect:

| Dialect | Training Wikipedia | Test Wikipedia | Test Web | Population % |
|---------|-------------------|----------------|---------|--------------|
| BA      | 100               | 36             | 18      | 8            |
| BE      | 100               | 77             | 38      | 17           |
| FR      | 50                | 7              | 3       | 2            |
| OS      | 100               | 64             | 32      | 14           |
| WS      | 100               | 7              | 3       | 2            |
| ZH      | 100               | 100            | 50      | 22           |

- Two more dialects have comparable Wikipedia coverage.
- These 6+2 dialects cover about 80% of Swiss German population.
- Average sentence length:
  - 18 words for Wikipedia
  - 15 words for Web
Outline

1. Introduction
2. Data
3. N-Gram model
4. Word-based model
   - The idea
   - Creating georeferenced rules
   - The identification process
   - Extensions
5. Conclusion
N-Gram model

Training:

- Create one language model for every dialect with Wikipedia training data
- SRILM toolkit
- Letter n-grams, not word n-grams!

Testing:

- Compute perplexity of every language model on the text
- Annotate the text with the dialect that obtained lowest perplexity (Biadsy et al. 2009)

Experiments:

- Experiments with 2-grams to 6-grams
- Best results with 5-grams, but small differences
N-Gram model

Training:
- Create one language model for every dialect with Wikipedia training data
- SRILM toolkit
- Letter n-grams, not word n-grams!

Testing:
- Compute perplexity of every language model on the text
- Annotate the text with the dialect that obtained lowest perplexity (Biadsy et al. 2009)

Experiments:
- Experiments with 2-grams to 6-grams
- Best results with 5-grams, but small differences
N-Gram model

Training:

- Create one language model for every dialect with Wikipedia training data
- SRILM toolkit
- Letter n-grams, not word n-grams!

Testing:

- Compute perplexity of every language model on the text
- Annotate the text with the dialect that obtained lowest perplexity (Biadsy et al. 2009)

Experiments:

- Experiments with 2-grams to 6-grams
- Best results with 5-grams, but small differences
### Results

| Dialect | **Wikipedia** | | | **Web** | | |
|---------|--------------|---|---|--------------|---|---|
|         | Precision    | Recall | F  | Precision    | Recall | F  |
| BA      | 34           | 61     | 44 | 27           | 61     | 37 |
| BE      | 78           | 51     | 61 | 51           | 47     | 49 |
| FR      | 28           | 71     | 40 | 10           | 33     | 15 |
| OS      | 63           | 64     | 64 | 50           | 38     | 43 |
| WS      | 58           | 100    | 74 | 14           | 33     | 20 |
| ZH      | 77           | 62     | 69 | 77           | 41     | 53 |
| **Weighted Average** | | | **62** | | | **46** |

- Percentage values of the 5-gram model.
- The average is weighted by the relative population sizes of the dialect regions.

16% performance drop from **Wikipedia** to **Web** data!
16% performance drop from Wikipedia to Web data!

Why?

- Only 1-2 authors per dialect on Wikipedia
- Authors use individual spelling rules
- Authors have specific topics of interest
  - ZH: Swiss politicians
  - OS: Religion
  - WS: Middle Ages
- The model learns to identify authors or topics rather than dialects.
- Instead of collecting more data, we tried another method.
Outline

1. Introduction
2. Data
3. N-Gram model
4. Word-based model
   - The idea
   - Creating georeferenced rules
   - The identification process
   - Extensions
5. Conclusion

Yves Scherrer: Word-based dialect identification with georeferenced rules
Outline

1. Introduction
2. Data
3. N-Gram model
4. Word-based model
   - The idea
   - Creating georeferenced rules
   - The identification process
   - Extensions
5. Conclusion
Word-based dialect identification

An old idea:

- Function words are very frequent and distinct enough to perform language identification. (Ingle 1980)

A very old idea:

- A *shibboleth* is “a use of language regarded as distinctive of a particular group.” (Merriam-Webster online dictionary)
  - “[The term] derives from an account in the Hebrew Bible, in which [the] pronunciation of this word was used to distinguish Ephraimites, whose dialect lacked a /ʃ/ sound, from Gileadites whose dialect did include such a sound.” (Wikipedia)

Our take:

- Almost every word is a potential shibboleth for some dialect.
- Instead of using function words only, we use the entire lexicon to perform dialect identification.
Word-based dialect identification

An old idea:

- Function words are very frequent and distinct enough to perform language identification. (Ingle 1980)

A very old idea:

- A *shibboleth* is “a use of language regarded as distinctive of a particular group.” (Merriam-Webster online dictionary)

  - “[The term] derives from an account in the Hebrew Bible, in which [the] pronunciation of this word was used to distinguish Ephraimites, whose dialect lacked a /ʃ/ sound, from Gileadites whose dialect did include such a sound.” (Wikipedia)

Our take:

- Almost every word is a potential shibboleth for some dialect.
- Instead of using function words only, we use the entire lexicon to perform dialect identification.
Word-based dialect identification

An old idea:

- Function words are very frequent and distinct enough to perform language identification. (Ingle 1980)

A very old idea:

- A shibboleth is “a use of language regarded as distinctive of a particular group.” (Merriam-Webster online dictionary)
  - “[The term] derives from an account in the Hebrew Bible, in which [the] pronunciation of this word was used to distinguish Ephraimites, whose dialect lacked a /ʃ/ sound, from Gileadites whose dialect did include such a sound.” (Wikipedia)

Our take:

- Almost every word is a potential shibboleth for some dialect.
- Instead of using function words only, we use the entire lexicon to perform dialect identification.
The lexicon

Requirement:

- A lexicon that relates each Swiss German word to its dialect(s), i.e. its area of occurrence.

Available:

- A Standard German word list with lemma, POS, morphology, and frequency information.
- A set of georeferenced transformation rules that translate Standard German words to Swiss German dialect words.
  - Originally intended for MT.
  - 300 phonetic rules (130 phenomena), 540 lexical rules (250 phenomena) and 130 morphological rules (60 phenomena).
  - Most rules are linked to probability maps extracted from the linguistic atlas SDS.
- Every Standard German word yields several Swiss German words by application of the transformation rules.
The lexicon

Requirement:

- A lexicon that relates each Swiss German word to its dialect(s), i.e. its area of occurrence.

Available:

- A Standard German word list with lemma, POS, morphology, and frequency information.

- A set of georeferenced transformation rules that translate Standard German words to Swiss German dialect words.
  
  - Originally intended for MT.
  - 300 phonetic rules (130 phenomena), 540 lexical rules (250 phenomena) and 130 morphological rules (60 phenomena).
  - Most rules are linked to probability maps extracted from the linguistic atlas SDS.

- Every Standard German word yields several Swiss German words by application of the transformation rules.
The lexicon

Requirement:

- A lexicon that relates each Swiss German word to its dialect(s), i.e. its area of occurrence.

Available:

- A Standard German word list with lemma, POS, morphology, and frequency information.
- A set of georeferenced transformation rules that translate Standard German words to Swiss German dialect words.
  - Originally intended for MT.
  - 300 phonetic rules (130 phenomena), 540 lexical rules (250 phenomena) and 130 morphological rules (60 phenomena).
  - Most rules are linked to probability maps extracted from the linguistic atlas SDS.
- Every Standard German word yields several Swiss German words by application of the transformation rules.
The lexicon

An example:

Standard German word: *suchen*-VVFIN-3.Pl.Pres.Ind

One possible derivation:
1. \( u \rightarrow ue \)
2. \( u \rightarrow u \) (not \( \ddot{u} \))
3. \( e \rightarrow a \) (in diphthong)
4. 3.Pl.Pres.Ind \( \rightarrow \) *end*

Resulting Swiss German word: *suachend*

Area of occurrence of this word: Created by pointwise product of the maps associated with rules 1-4 (joint probability).
The lexicon

Requirement:

- A lexicon that relates each Swiss German word to its dialect(s), i.e. its area of occurrence.

Available:

- A Standard German word list with lemma, POS, morphology, and frequency information.
- A set of georeferenced transformation rules that translate Standard German words to Swiss German dialect words.
  - Originally intended for MT.
  - 300 phonetic rules (130 phenomena), 540 lexical rules (250 phenomena) and 130 morphological rules (60 phenomena).
  - Most rules are linked to probability maps extracted from the linguistic atlas SDS.
- Every Standard German word yields several Swiss German words by application of the transformation rules.
Outline

1. Introduction
2. Data
3. N-Gram model
4. Word-based model
   - The idea
   - Creating georeferenced rules
   - The identification process
   - Extensions
5. Conclusion
Georeferenced rules

Original SDS map (SDS II/120):
Georeferenced rules

Map processing:

- The map exemplifies the outcome of final *nd* with the word *Hund*.
- Four main variants: *nd* [nd̥], *nt* [nt], *ng* [ŋ], *nn* [nː].
- Two minor variants of *ng* and *nn*.

Processing assumptions:

- The same map applies to all/most other words with final *-nd*.
- The two minor variants are not distinguished by dialect writers and by the spelling conventions we adopt (Dieth 1986).
- The pronunciations recorded at the inquiry points allow to interpolate the pronunciations at other locations.
  → Digitize and interpolate the map.
Georeferenced rules

Map processing:

- The map exemplifies the outcome of final *nd* with the word *Hund*.
- Four main variants: *nd* [n̥d], *nt* [nt], *ng* [ŋ], *nn* [nː].
- Two minor variants of *ng* and *nn*.

Processing assumptions:

- The same map applies to all/most other words with final -*nd*.
- The two minor variants are not distinguished by dialect writers and by the spelling conventions we adopt (Dieth 1986).
- The pronunciations recorded at the inquiry points allow to interpolate the pronunciations at other locations.
  → Digitize and interpolate the map.
Georeferenced rules

Digitized point map:

nd  ng  nt  nn
Georeferenced rules

Set of interpolated surface maps:

Algorithm: Kernel density estimator (Rumpf et al. 2009)
Normalization: The shade of the map represents the probability of the variant (black: $p = 1$, white: $p = 0$). At each geographic point, the probabilities of all variants of a phenomenon sum up to 1.
A georeferenced transfer rule refers to a rewriting rule valid in a particular geographic region.

One linguistic phenomenon usually consists of several, geographically complementary transfer rules.

- Rules are implemented with regular expressions.

Example of phonetic rules:

- $-nd \rightarrow -nn$
- $-nd \rightarrow -ng$
- $-nd \rightarrow -nt$
- $-nd \rightarrow -nd$
Phonetic rules cannot generate all dialect words.

Example of lexical rules:

- \textit{immer} $\rightarrow$ geng
- \textit{immer} $\rightarrow$ \textit{immer}
- \textit{immer} $\rightarrow$ all

Example morphological rules:

1.\textit{Pl.Pres} $\rightarrow$ -e
2.\textit{Pl.Pres} $\rightarrow$ -ed
3.\textit{Pl.Pres} $\rightarrow$ -id
4.\textit{Pl.Pres} $\rightarrow$ -end
The lexicon

Requirement:

- A lexicon that relates each Swiss German word to its dialect(s), i.e. its area of occurrence.

Available:

- A Standard German word list with lemma, POS, morphology, and frequency information.

- A set of georeferenced transformation rules that translate Standard German words to Swiss German dialect words.
  
  - Originally intended for MT.
  - 300 phonetic rules (130 phenomena), 540 lexical rules (250 phenomena) and 130 morphological rules (60 phenomena).
  - Most rules are linked to probability maps extracted from the linguistic atlas SDS.

- Every Standard German word yields several Swiss German words by application of the transformation rules.
Segment identification

Unidentified input segment (bag of words).

Assign a map to every derivation of a word.

Lexicon lookup (pointwise product of rule maps).

Pointwise maximum.

Combined derivation maps into word maps.

Combined word maps into a sentence map.

Combine word maps into a sentence map.

Pointwise product (joint probability).

Colored ellipses represent maps.
Segment identification

Unidentified input segment (bag of words).

Assign a map to every **derivation** of a word.

Lexicon lookup (pointwise product of rule maps).

Combine derivation maps into word maps.

Pointwise maximum.

Combine word maps into a sentence map.

Pointwise product (joint probability).

Colored ellipses represent maps.
Segment identification

Unidentified input segment (bag of words).

Assign a map to every derivation of a word.

Lexicon lookup (pointwise product of rule maps).

Combine derivation maps into word maps.

Pointwise maximum.

Combine word maps into a sentence map.

Pointwise product (joint probability).

Colored ellipses represent maps.
Segment identification

Unidentified input segment (bag of words).

Assign a map to every **derivation** of a word.

Combine derivation maps into **word** maps.

Combine word maps into a **sentence** map.

Lexicon lookup (pointwise product of rule maps).

Pointwise maximum.

Pointwise product (joint probability).

Colored ellipses represent maps.
Derivation weighting I

What if a text contains the word *dr*?

- **der-ART → dr**
  Derivation valid only in Western dialects.
  7800 occurrences of *der* in Standard German corpus.

- **Dr.-NN → dr**
  Derivation valid in all dialects.
  11 occurrences of *Dr.* in Standard German corpus.

What can we infer from the occurrence of *dr*?

- The text could be written in any dialect.

  More likely to occur in the West, because *der-ART* is much more frequent than *Dr.-NN*.

Weight each derivation by the **frequency of the Standard German word** (the one at the start of the derivation).
Derivation weighting I

What if a text contains the word *dr*?

- **der-ART → dr**
  Derivation valid only in Western dialects.
  7800 occurrences of *der* in Standard German corpus.

- **Dr.-NN → dr**
  Derivation valid in all dialects.
  11 occurrences of *Dr.* in Standard German corpus.

What can we infer from the occurrence of *dr*?

- The text could be written in any dialect.

- More likely to occur in the West, because *der*-ART is much more frequent than *Dr.*-NN.

Weight each derivation by the **frequency of the Standard German word** (the one at the start of the derivation).
We only want to include shibboleths in our model, i.e. derivations with a high discriminative potential.

- A derivation that is valid in 95% of German-speaking Switzerland has very low discriminative potential.
- It is most likely not a shibboleth, and its usefulness for dialect ID is limited.

Weight each derivation by its **discriminative potential**: The discriminative potential is the inverse of the cumulated probability of the derivation.

**Combination:**
Weight each derivation by both its frequency and its discriminative potential.
Derivation weighting II

We only want to include shibboleths in our model, i.e. derivations with a high discriminative potential.

- A derivation that is valid in 95% of German-speaking Switzerland has very low discriminative potential.
- It is most likely not a shibboleth, and its usefulness for dialect ID is limited.

Weight each derivation by its **discriminative potential**: The discriminative potential is the inverse of the cumulated probability of the derivation.

**Combination:**
Weight each derivation by both its frequency and its discriminative potential.
## Results

### No weighting:

| Dialect | *Wikipedia* F-Measure | *Web* F-Measure |
|---------|-----------------------|-----------------|
| BA      | 26                    | 35              |
| BE      | 50                    | 59              |
| FR      | 0                     | 22              |
| OS      | 38                    | 43              |
| WS      | 5                     | 13              |
| ZH      | 44                    | 46              |
| **W. Avg.** | **40**                | **46**          |

### Discriminative potential weighting:

| Dialect | *Wikipedia* F-Measure | *Web* F-Measure |
|---------|-----------------------|-----------------|
| BA      | 32                    | 23              |
| BE      | 47                    | 63              |
| FR      | 13                    | 25              |
| OS      | 40                    | 56              |
| WS      | 7                     | 0               |
| ZH      | 34                    | 55              |
| **W. Avg.** | **37**                | **51**          |

### Frequency weighting:

| Dialect | *Wikipedia* F-Measure | *Web* F-Measure |
|---------|-----------------------|-----------------|
| BA      | 40                    | 32              |
| BE      | 53                    | 68              |
| FR      | 0                     | 0               |
| OS      | 30                    | 48              |
| WS      | 15                    | 22              |
| ZH      | 53                    | 58              |
| **W. Avg.** | **44**                | **53**          |

### Frequency and discriminative potential:

| Dialect | *Wikipedia* F-Measure | *Web* F-Measure |
|---------|-----------------------|-----------------|
| BA      | 35                    | 17              |
| BE      | 54                    | 69              |
| FR      | 0                     | 25              |
| OS      | 33                    | 47              |
| WS      | 13                    | 20              |
| ZH      | 58                    | 58              |
| **W. Avg.** | **46**                | **52**          |
Discussion

Impact of weighting techniques:
- Discriminative potential weighting only slightly improves performance on the web corpus.
- Frequency weighting helps on both corpora.
- The two techniques are additive.

The model performs better on Web data than on Wikipedia.
- Spelling? Standard German influence on Wikipedia (translation)? Proper nouns?

Upper bound?
- Not every sentence can be assigned unambiguously to one dialect.

Model comparison:
- N-gram model outperforms word-based model on Wikipedia (overfitting).
- Word-based model outperforms n-gram model on Web.
Discussion

Impact of weighting techniques:
- Discriminative potential weighting only slightly improves performance on the web corpus.
- Frequency weighting helps on both corpora.
- The two techniques are additive.

The model performs better on Web data than on Wikipedia.
- Spelling? Standard German influence on Wikipedia (translation)? Proper nouns?

Upper bound?
- Not every sentence can be assigned unambiguously to one dialect.

Model comparison:
- N-gram model outperforms word-based model on Wikipedia (overfitting).
- Word-based model outperforms n-gram model on Web.
Impact of weighting techniques:
- Discriminative potential weighting only slightly improves performance on the web corpus.
- Frequency weighting helps on both corpora.
- The two techniques are additive.

The model performs better on Web data than on Wikipedia.
- Spelling? Standard German influence on Wikipedia (translation)? Proper nouns?

Upper bound?
- Not every sentence can be assigned unambiguously to one dialect.

Model comparison:
- N-gram model outperforms word-based model on Wikipedia (overfitting).
- Word-based model outperforms n-gram model on Web.
Discussion

Impact of weighting techniques:
- Discriminative potential weighting only slightly improves performance on the web corpus.
- Frequency weighting helps on both corpora.
- The two techniques are additive.

The model performs better on **Web** data than on **Wikipedia**.
- Spelling? Standard German influence on **Wikipedia** (translation)? Proper nouns?

Upper bound?
- Not every sentence can be assigned unambiguously to one dialect.

Model comparison:
- N-gram model outperforms word-based model on **Wikipedia** (overfitting).
- Word-based model outperforms n-gram model on **Web**.
Outline

1 Introduction
2 Data
3 N-Gram model
4 Word-based model
   - The idea
   - Creating georeferenced rules
   - The identification process
   - Extensions
5 Conclusion
Evaluation

Result of our model:

Max  It is BE because the peak probability of the BE region is higher than the peak probability of any other region (results above).

Avg  It is BE because the average probability of the BE region is higher than the average probability of any other region.

Which one is it?

Yves Scherrer: Word-based dialect identification with georeferenced rules
Extensions

Large heterogeneous regions:

- **Max** works better:
  The peak is most likely inside the region.

- Heterogeneity drags down the **Avg** value.

Small homogeneous regions:

- **Avg** works better:
  Homogeneity keeps average value high.

- The peak used by **Max** might be slightly outside the region.
Extensions

**Max**  Bias towards large, heterogeneous regions.

**Avg**  Bias towards small, homogeneous regions.

**Cmb**  Use area size to determine the better performing measure:

- Max for regions > 5% of Swiss territory, Avg otherwise.

**Ora**  Use an oracle to determine the better performing measure.

| Dialect | Wikipedia  | | | | | Web  | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Cmb | Ora | Max | Avg | Cmb | Ora | Max | Avg | Cmb | Ora |
| BA | 35 | 32 | 32 | 35 | 17 | 43 | 43 | 43 | 17 | 43 | 43 | 43 |
| BE | 54 | 39 | 54 | 54 | 69 | 54 | 69 | 69 | 69 | 54 | 69 | 69 | 69 |
| FR | 0 | 7 | 7 | 7 | 25 | 11 | 11 | 25 | 25 | 11 | 11 | 25 |
| OS | 33 | 23 | 33 | 33 | 47 | 49 | 47 | 49 | 47 | 49 | 47 | 49 |
| WS | 13 | 13 | 13 | 13 | 20 | 31 | 20 | 31 | 20 | 31 | 20 | 31 |
| ZH | 58 | 60 | 60 | 60 | 58 | 68 | 68 | 68 | 58 | 68 | 68 | 68 |
| W. Avg. | 46 | 40 | 47 | 47 | 52 | 55 | 58 | 59 | 52 | 55 | 58 | 59 |
Topics of further investigation

Improve rule coverage
- About 35% of words are not in lexicon
- Clitics
- Most obvious cases of orthographic variation

Assess the impact of lexicon size on ID accuracy
- Zipf's law suggests that a smaller lexicon with high-frequency entries may be sufficient.

It is hard to identify the dialect of single sentences
- Results on entire paragraphs (of various lengths) are 20% higher (absolute) on average.
Conclusion

Comparison of a character n-gram model with a word-based model for dialect identification

- N-gram model overfits: training corpora are too small and topic-dependent.
- Word-based model relies on external resources, thus more robust.

Word-based model uses continuous maps

- Does not require a priori dialect classification.
- Dialect classes are only introduced for evaluation purposes.

Reuse of existing dialectological resources

- Machine translation, dialect identification, tagging, parsing, dialectometric studies, ...

Details in our EMNLP paper!
Conclusion

Comparison of a character n-gram model with a word-based model for dialect identification

- N-gram model overfits: training corpora are too small and topic-dependent.
- Word-based model relies on external resources, thus more robust.

Word-based model uses continuous maps

- Does not require a priori dialect classification.
- Dialect classes are only introduced for evaluation purposes.

Reuse of existing dialectological resources

- Machine translation, dialect identification, tagging, parsing, dialectometric studies, ...

Details in our EMNLP paper!
Conclusion

Comparison of a character n-gram model with a word-based model for dialect identification

- N-gram model overfits: training corpora are too small and topic-dependent.
- Word-based model relies on external resources, thus more robust.

Word-based model uses continuous maps

- Does not require a priori dialect classification.
- Dialect classes are only introduced for evaluation purposes.

Reuse of existing dialectological resources

- Machine translation, dialect identification, tagging, parsing, dialectometric studies,

Details in our EMNLP paper!
Conclusion

Comparison of a character n-gram model with a word-based model for dialect identification

- N-gram model overfits: training corpora are too small and topic-dependent.
- Word-based model relies on external resources, thus more robust.

Word-based model uses continuous maps

- Does not require a priori dialect classification.
- Dialect classes are only introduced for evaluation purposes.

Reuse of existing dialectological resources

- Machine translation, dialect identification, tagging, parsing, dialectometric studies, ...

Details in our EMNLP paper!
