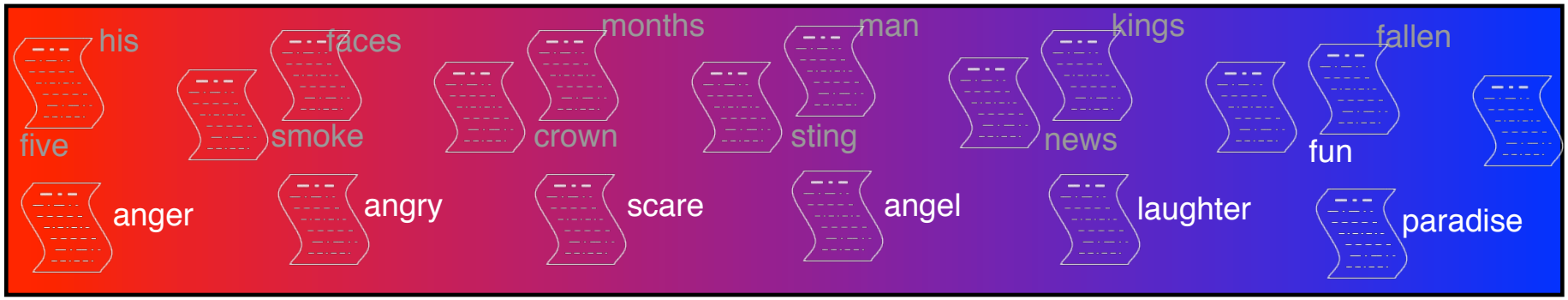
Multilingual Sentiment Analysis Using Latent Semantic Indexing and Machine Learning



Brett Bader, Digital Globe, bbader@digitalglobe.com

Philip Kegelmeyer, Sandia National Laboratories, wpk@sandia.gov

Peter Chew, Galisteo Consulting Group Inc, PeterAChew@aol.com

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

SENTIRE, December 11, 2011

Overview

We treat **multilingual document sentiment classification**

as a supervised machine learning problem, which requires:

- multilingual document attributes
- monolingual ground truth documents

We assess performance,

raise an objection,

and address it.

Document Sentiment Classification

Psalm 126:2–4

Then was our mouth filled with laughter, and our tongue with singing: then said they among the heathen, The LORD hath done great things for them. The LORD hath done great things for us; whereof we are glad. Turn again our captivity, O LORD, as the streams in the south.

Revelation 9:18–19

By these three was the third part of men killed, by the fire, and by the smoke, and by the brimstone, which issued out of their mouths. For their power is in their mouth, and in their tails: for their tails were like unto serpents, and had heads, and with them they do hurt.

Document Sentiment Classification, Multilingually

Salmos 126:2–4

Entonces nuestra boca se henchirá de risa, Y nuestra lengua de alabanza; Entonces dirán entre las gentes: Grandes cosas ha hecho Jehová con éstos. Grandes cosas ha hecho Jehová con nosotros; Estaremos alegres. Haz volver nuestra cautividad oh Jehová, Como los arroyos en el austro.

Apocalipsis 9:18–19

De estas tres plagas fué muerta la tercera parte de los hombres: del fuego, y del humo, y del azufre, que salan de la boca de ellos. Porque su poder está en su boca y en sus colas: porque sus colas eran semejantes serpientes, y tenían cabezas, y con ellas dañan.

Overview

We treat multilingual document sentiment classification

as a **supervised machine learning** problem, which requires:

- multilingual document attributes
- monolingual ground truth documents

We assess performance,

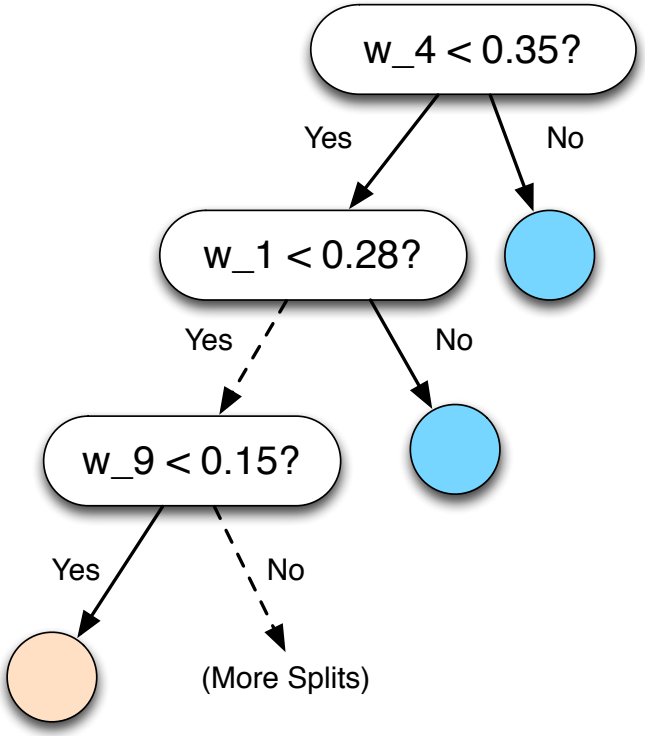
raise an objection,

and address it.

Supervised Machine Learning Overview

Also known as: pattern recognition, statistical inference, data mining.

- Input: "ground truth" data.
  - Samples, with attributes and labels.
  - For document sentiment analysis:
    - Samples: documents
    - Attributes: concept weights
    - Labels: positive, negative
- Apply suitable method:
  decision trees, neural nets, SVMs.
- Output:
  rules for labeling new, *unlabeled* documents.



Decision tree representation.

Overview

We treat multilingual document sentiment classification

as a supervised machine learning problem, which requires:

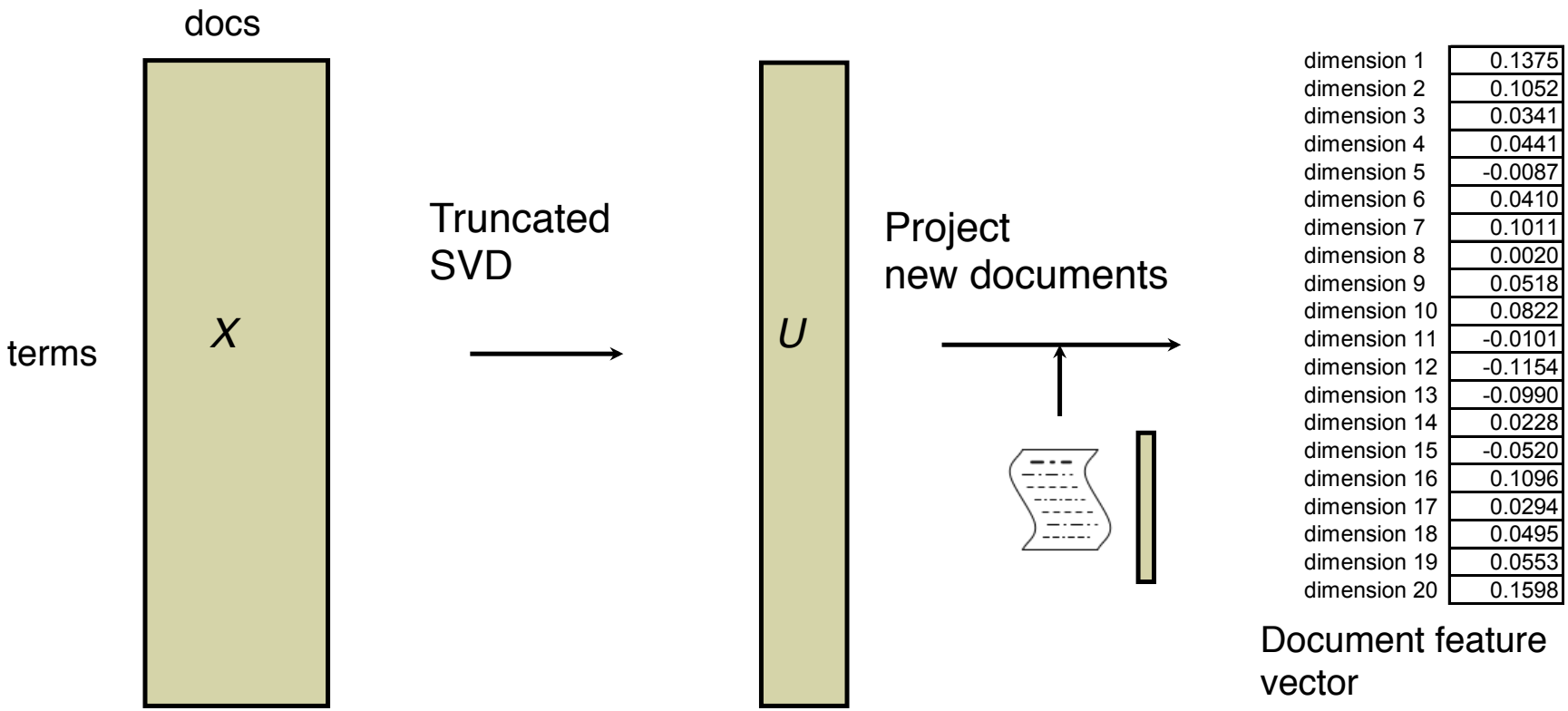
- **multilingual document attributes**
- monolingual ground truth documents

We assess performance,

raise an objection,

and address it.

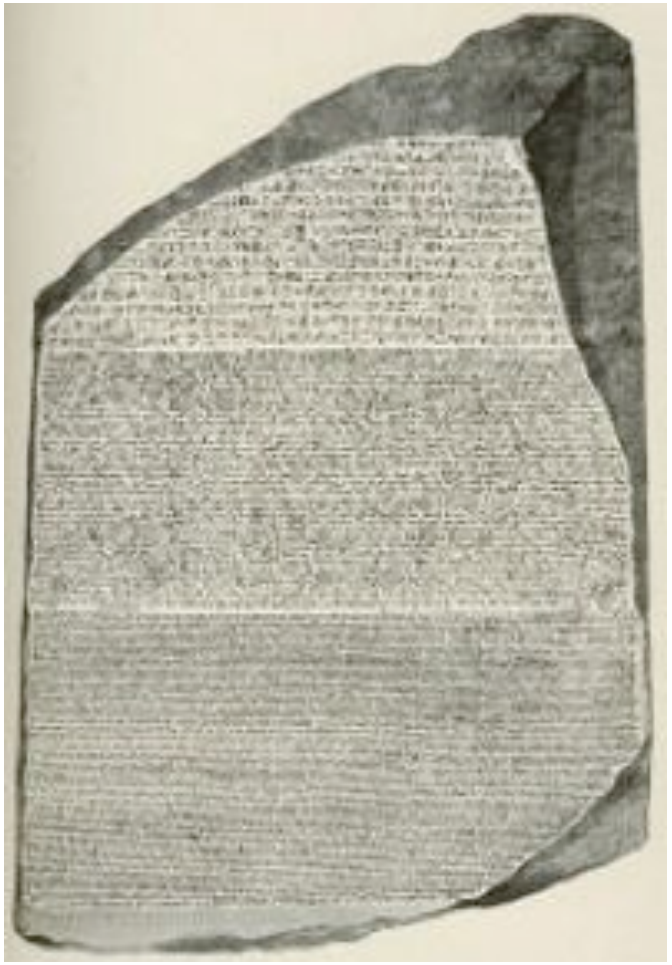
Monolingual Latent Semantic Indexing



Result: documents represented by short set of "concept weights"[1].

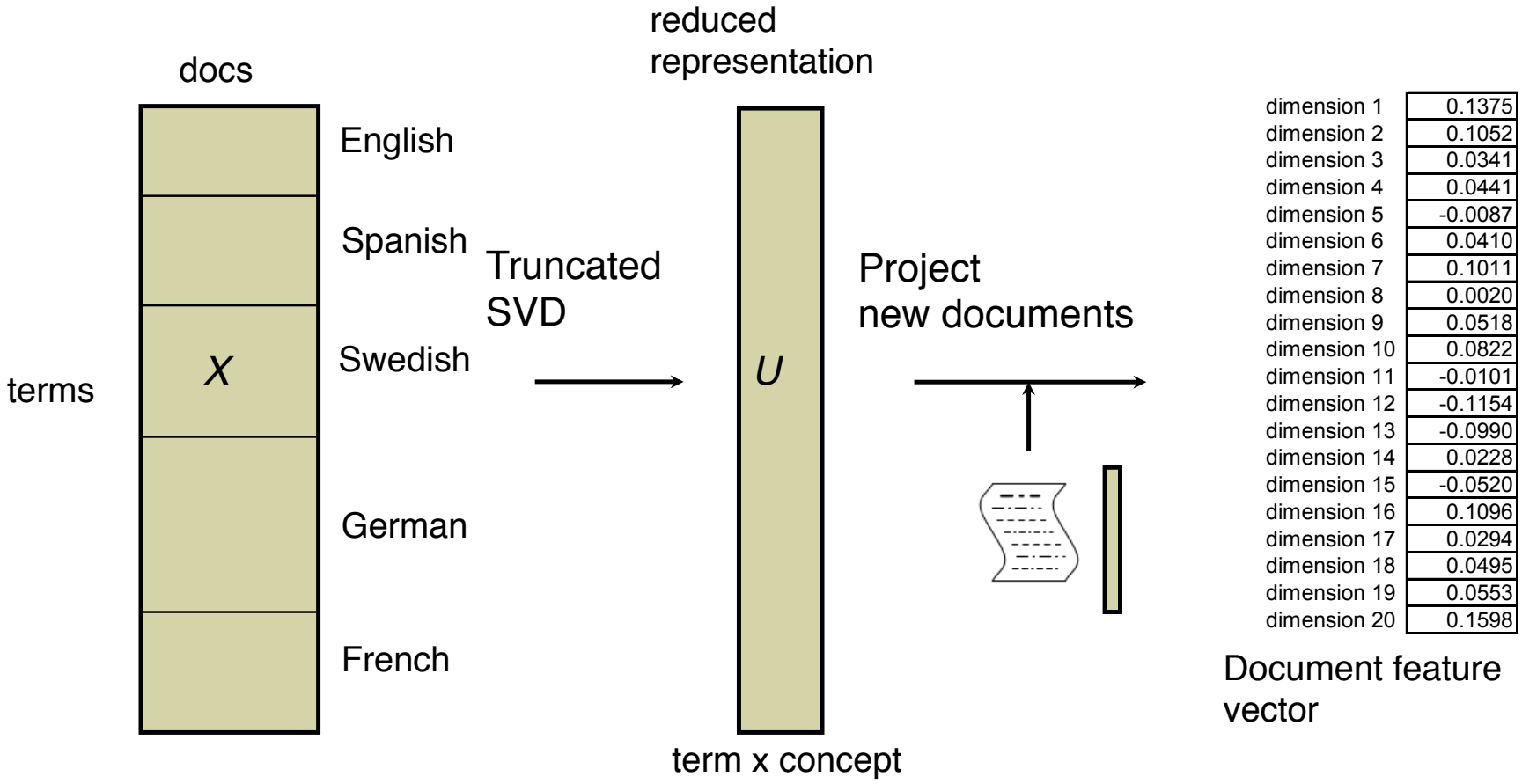
Europarl Corpus as Our "Rosetta Stone"

- Translations of the proceedings of the European Parliament.
- Sentence aligned text
  16M sentences in 11 languages.
- 1,247,832 speeches.
- 1,249,253 terms in 11 languages.



The Rosetta Stone

Multilingual Latent Semantic Indexing



Result: documents represented by *language-independent* features[2, 3].

Overview

We treat multilingual document sentiment classification

as a supervised machine learning problem, which requires:

- multilingual document attributes
- **monolingual ground truth documents**

We assess performance,

raise an objection,

and address it.

Groundtruthing the Bible for Sentiment

- Could use exhaustive human reading and judgment. (Requires exhaustion. And judgment).
- We used a sentiment lexicon to bootstrap the process. (Sentiment lexicon not strictly required; any accurate labeling mechanism suffices.)
- A sentiment lexicon[4] maps terms to "valence"

Term	Valence, 0 to 9
ace	6.88
ache	2.26
...	...
fun	8.27
funeral	1.39

...
Chronicles 4
Chronicles 5
Chronicles 6
...
Psalm 125
Psalm 126
Psalm 127
...
Revelation 8
Revelation 9
Revelation 10
Revelation 11
...

Initial Scoring for Each Bible Chapter

- For each chapter
  - Add up (centered) valences
  - Normalize by number of terms
- Find the 100 most positive, 100 most negative.
- Inspect only those, to confirm.

Term	Valence
In the beginning God	8.15
created the heavens	7.30
and the earth .	7.15
And the earth	7.15
was waste	2.93
and void. . .	

(Genesis, Chapter 1, ranks 227 out of 1188 chapters.)

Hand Inspection Was Necessary

Revelation 9:1–12 (a demonic plague of locusts) scored *positive.*

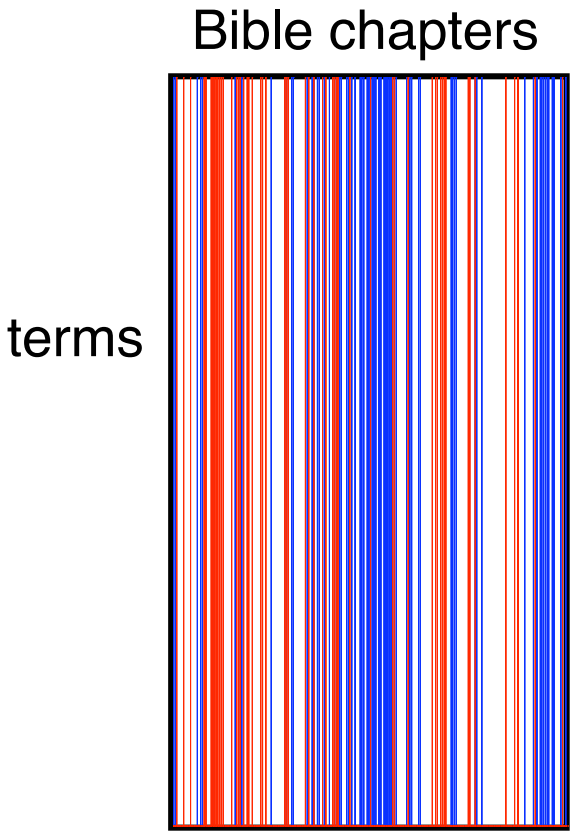
1 The fifth angel sounded his trumpet, and I saw a star that had fallen from the sky to the earth.
The star was given the key to the shaft of the Abyss. 2 When he opened the Abyss, smoke rose from it
like the smoke from a gigantic furnace. The sun and sky were darkened by the smoke from the Abyss. 3
And out of the smoke locusts came down upon the earth and were given power like that of scorpions of
the earth. 4 They were told not to harm the grass of the earth or any plant or tree, but only those
people who did not have the seal of God on their foreheads. 5 They were not given power to kill them,
but only to torture them for five months. And the agony they suffered was like that of the sting of a
scorpion when it strikes a man. 6 During those days men will seek death, but will not find it; they will long
to die, but death will elude them.
7 The locusts looked like horses prepared for battle. On their heads they wore something like crowns of
gold, and their faces resembled human faces. 8 Their hair was like women's hair, and their teeth were like
lions' teeth. 9 They had breastplates like breastplates of iron, and the sound of their wings was like the
thundering of many horses and chariots rushing into battle. 10 They had tails and stings like scorpions, and
in their tails they had power to torment people for five months. 11 They had as king over them the
angel of the Abyss, whose name in Hebrew is Abaddon, and in Greek, Apollyon.

12 The first woe is past; two other woes are yet to come. . . .

Why? Lexicon lacked "smoke", "Abyss", "sting", "scorpion", . . .

Final Sentiment Groundtruth Dataset

- Manual inspection turned up a few "Revelations 9" problems.
- Weeded by hand, and re-seeded.
- Final result (out of 1188 chapters)
  - 115 positive chapters
  - 78 negative chapters
  - 59.6% positive



Red, Blue, White for
Positive, Negative, Unlabeled

Overview

We treat multilingual document sentiment classification

as a supervised machine learning problem, which requires:

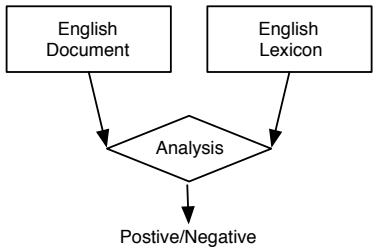
- multilingual document attributes
- monolingual ground truth documents

We **assess performance**,

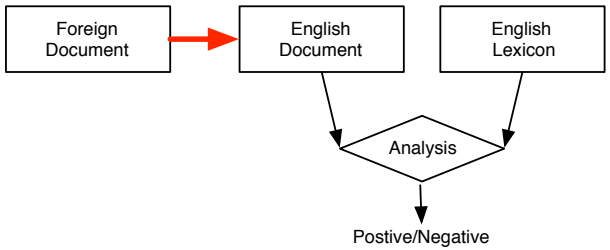
raise an objection,

and address it.

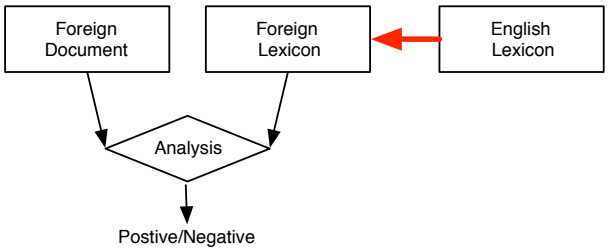
Pause: What Have Other People Done?



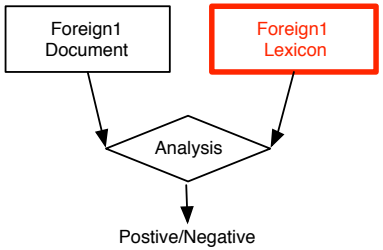
Monolingual

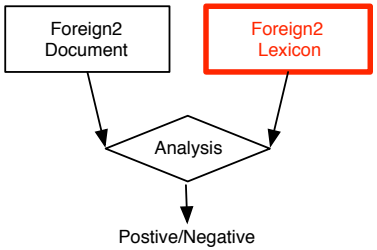


Foreign Document to English[5]

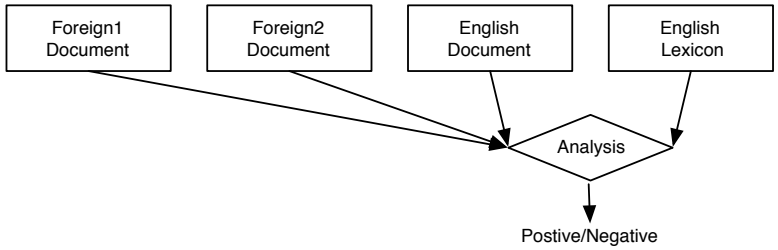


English Lexicon to Foreign[6]





Foreign Lexicon Replication[7]

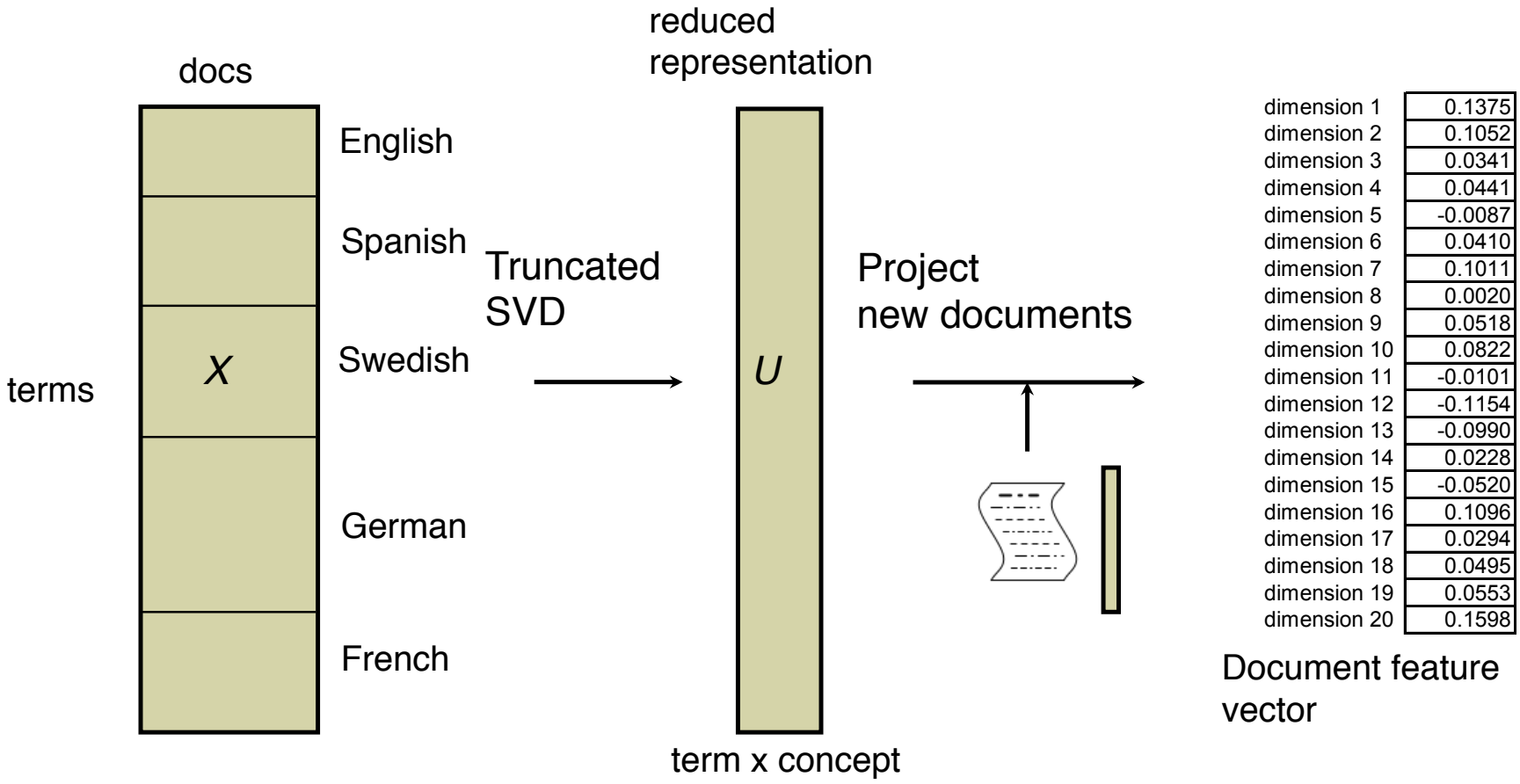


LSI plus Machine Learning

Our method requires no translation.

Further, it needs a sentiment lexicon only as one way to boot strap.

Process English Bible Chapters through Europarl SVD



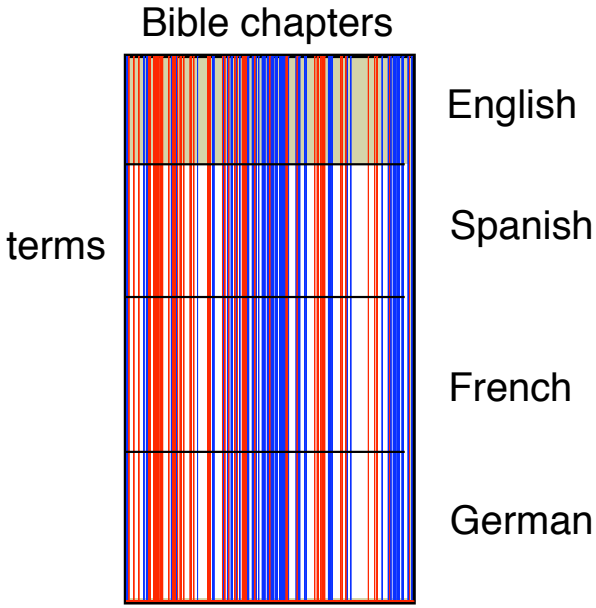
Use only the Labeled Chapters . . .

. . . to generate training data that looks like:

Chapter	Valence	w_1	w_2	w_3	. . .	w_{300}
c_1, Psalm 126	Positive	0.12	0.03	0.97	. . .	0.12
c_2, Psalm 127	Positive	0.99	0.02	0.33	. . .	0.03
c_3, Chronicles 5	Negative	0.30	0.27	0.12	. . .	0.13
c_4, Revelation 10	Positive	0.16	0.83	0.08	. . .	0.58
c_5, Chronicles 5	Negative	0.17	0.65	0.36	. . .	0.64
c_6, Ezra 10	Negative	0.44	0.12	0.29	. . .	0.42
c_7, Ezekiel 5	Negative	0.42	0.24	0.33	. . .	0.88
c_8, James3	Positive	0.78	0.42	0.44	. . .	0.52
⋮	⋮	⋮	⋮	⋮		⋮
c_{193}, Revelation 9	Negative	0.12	0.41	0.92	. . .	0.17

Test on the Foreign, Labeled Chapters

- Build an ensemble of bagged decision trees.
- Process foreign language *test* chapters through Europarl SVD.
- Assume that sentiment is preserved across languages.
- Use the ensemble to classify the 3x193 chapters in Spanish, French, German.
- Result:
  - Accuracy of 74.9%
  - Statistically significantly (one-sample t-test, $\alpha = 0.01$) better than the ...
  - Baseline random accuracy of 56.9%



Overview

We treat multilingual document sentiment classification

as a supervised machine learning problem, which requires:

- multilingual document attributes
- monolingual ground truth documents

We assess performance,

raise an objection,

and address it.

Have We Simply Learned Topic, Not Sentiment?

Maybe. So shuffle verses within same-sentiment chapters.

A toy example, from the positive chapters. Before:

Chapter/Verse	Topic	Excerpt
Psalms 126, Verse 2	rejoicing	"... laughter ..."
Psalms 126, Verse 3	rejoicing	"... joy ..."
Psalms 126, Verse 4	rejoicing	"... fortunes ..."
Psalms 126, Verse 5	rejoicing	"... songs ..."
Revelation 10, Verse 2	prophets	"... book ..."
Revelation 10, Verse 4	prophets	"... write ..."
Revelation 10, Verse 11	prophets	"... prophesy ..."
James 3, Verse 1	wisdom	"... teachers ..."
James 3, Verse 13	wisdom	"... wisdom ..."

Have We Simply Learned Topic, Not Sentiment?

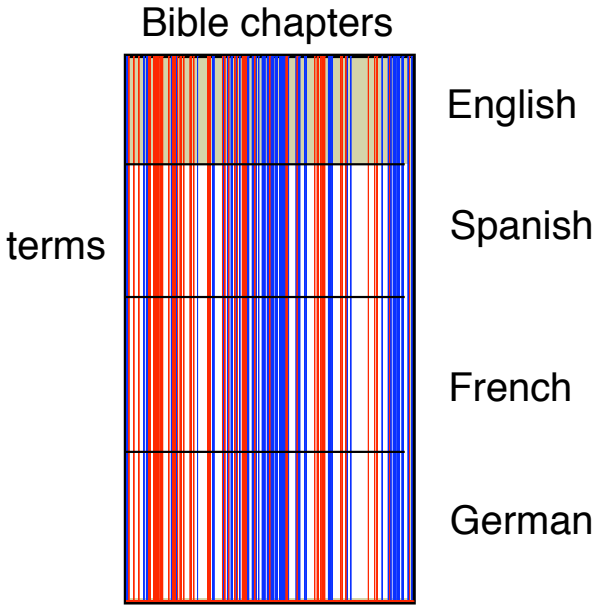
And after:

Chapter/Verse	Topic	Excerpt
Psalms 126, Verse 4	rejoicing	"... fortunes ..."
James 3, Verse 13	wisdom	"... wisdom ..."
Revelation 10, Verse 4	prophets	"... write ..."
Psalms 126, Verse 5	rejoicing	"... songs ..."
Revelation 10, Verse 11	prophets	"... prophesy ..."
James 3, Verse 1	wisdom	"... teachers ..."
Psalms 126, Verse 3	rejoicing	"... joy ..."
Revelation 10, Verse 2	prophets	"... book ..."
Psalms 126, Verse 2	rejoicing	"... laughter ..."

Break up positive topics by re-distributing their sentences.
Do the same, separately, with negative topics.

Train on Shuffled English, Test on Foreign

- Process shuffled chapters through SVD.
- Generates new, topic-incoherent, training data.
- Train a new ensemble from the new training data.
- Use the new ensemble to classify the 3x193 chapters in Spanish, French, German.
- Result:
  - Accuracy of 72.0%
  - Still significantly better than the 56.9% baseline.
  - But lower than 74.9%.
- Indicates that some, but not all, of sentiment is bound up in topic.



Conclusion

- We have a demonstrated a supervised machine learning approach to determine sentiment in multilingual documents.
  - Does not require translation
  - Uses a sentiment lexicon only for bootstrapping sentiment labels
  - Uses LSA to project documents into a language-independent space.
  - Uses machine learning on these features to build a predictive model

  Extensions:
  - Could easily be used with other topic models, such as LDA or NMF.
  - Could be applied to other emotional dimensions or meta-properties, such as "framing language"; prior similar application has been seen in characterizing ideology[8] in multilingual text.

References

[1] S. Deerwester, "Improving Information Retrieval with Latent Semantic Indexing," in *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, C. L. Borgman and E. Y. H. Pai, Eds., vol. 25. Atlanta, Georgia: American Society for Information Science, Oct. 1988.

[2] P. A. Chew, B. W. Bader, S. Helmreich, A. Abdelali, and S. J. Verzi, "An information-theoretic, vector-space model approach to cross-language information retrieval," *Journal of Natural Language Engineering*, 2010.

[3] B. Bader and P. Chew, *Text Mining: Applications and Theory*. Wiley, 2010, ch. Algebraic Techniques for Multilingual Document Clustering.

[4] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," *Technical Report C-1, The Center for Research in Psychophysiology University*, 1999.

[5] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *ICDE Workshops*. IEEE Computer Society, 2008, pp. 507–512.

[6] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 976–983.

[7] A.-L. Ginsca, E. Boros, A. Iftene, D. Trandabat, M. Toader, M. Corici, C.-A. Perez, and D. Cristea, "Sentimatrix — multilingual sentiment analysis service," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 189–195.

[8] P. Chew, P. Kegelmeyer, B. Bader, and A. Abdelali, "The knowledge of good and evil: Multilingual ideology classification with PARAFAC2 and machine learning," *Language Forum*, vol. 34, no. 1, pp. 37–52, 2008.

[9] B. Bader and P. Chew, "Enhancing multilingual latent semantic analysis with term alignment information," in *COLING 2008*, 2008.

[10] P. Chew and A. Abdelali, "Benefits of the massively parallel Rosetta Stone: Cross-language information retrieval with over 30 languages," in *Proceedings of the Association for Computational Linguistics*, 2007, pp. 872–879.

Philip Kegelmeyer, wpk@sandia.gov, `csmr.sandia.gov/~wpk`