Using a Bilingual Dictionary to Expand Topic Model Training Data

Kevin Prehn and Dr. John Jeffrey

Elmhurst College

NCUR

4/11/19

Outline

- Background
  - Semantic Similarity
  - Low Resource Languages
  - Test Problem
- Method
  - **Training data expansion**
- Results
- Discussion
- Future Work
- Questions

Background

Semantic similarity problem

- Given two documents, how **similar** are they in **meaning**?
- Corpus based methods [1, 2]
  - Large amount of data
- Knowledge based methods [1]
  - Semantic networks (Wordsnet)
- Can be framed as a **classification problem**

Low Resource Languages

- Not enough data for Corpus Based methods
- Not enough resources for Knowledge based methods
- A Bilingual dictionary is an early resource

RQ1: How might a Bilingual Dictionary be used to improve training data for a model for a Low Resource Language?

RQ2: How does using training data modified with a Bilingual Dictionary affect a model's performance compared to models trained with unmodified training data?

Test Problem Overview

- Topic Classification
- Training data
  - 4 Topic Models
    - Wikipedia articles [3, 4]
  - Background Model
    - Google Data [5]
- Testing data
  - NYT News Articles [6]
- Bilingual Dictionary
  - English-German dict.cc [7]

The New York Times

dict.cc
Deutsch-Englisch-Wörterbuch

Text Classification Method

- Probabilistic Latent Semantic Analysis (PLSA) [8]
- Training data calculates Language/Topic Models (thetas)

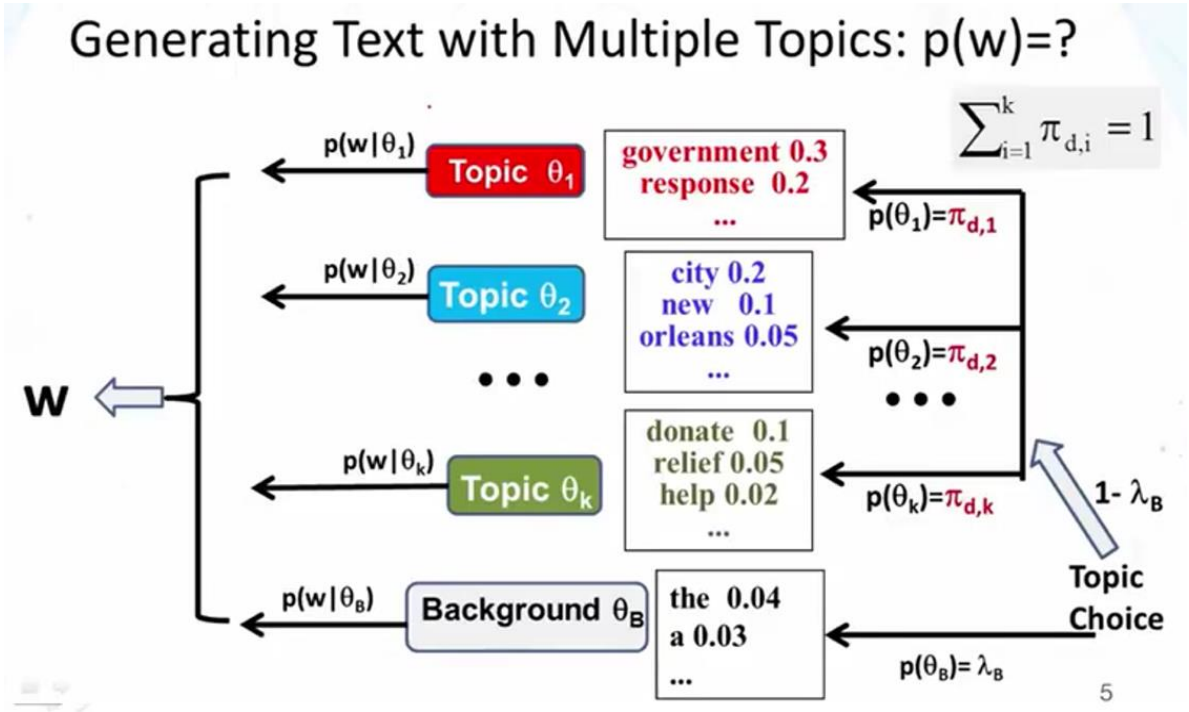


Diagram from [3]

Text Classification Method

– E-step $p(z_{d,w} = j) \propto \pi_{d,j}^{(n)} p^{(n)}(w \mid \theta_j)$ $\sum_{j=1}^{k} p(z_{d,w} = j) = 1$

$p(z_{d,w} = B) \propto \lambda_B p(w \mid \theta_B)$

– M-step

$\pi_{d,j}^{(n+1)} \propto \sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)$ $\forall d \in C, \sum_{j=1}^{k} \pi_{d,j} = 1$

$p^{(n+1)}(w \mid \theta_j) \propto \sum_{d \in C} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)$ $\forall j \in [1,k], \sum_{w \in V} p(w \mid \theta_j) = 1$

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{j'} \sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j')}$$

$$p^{(n+1)}(w \mid \theta_j) = \frac{\sum_{d \in C} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w',d)(1 - p(z_{d,w'} = B)) p(z_{d,w'} = j)}$$

Diagram from [3]

Method

Normal/Control Training Data

Collect Data

- 1 document per topic

Clean Data

- Remove punctuation, etc.

Calculate Models

- Word frequency

Number of Terms in Model by Topic			
Business	Politics	Science	Sports
819	978	2,356	1,258

Expanded Training Data (**RQ1**)

Collect Data
- 1 document per topic

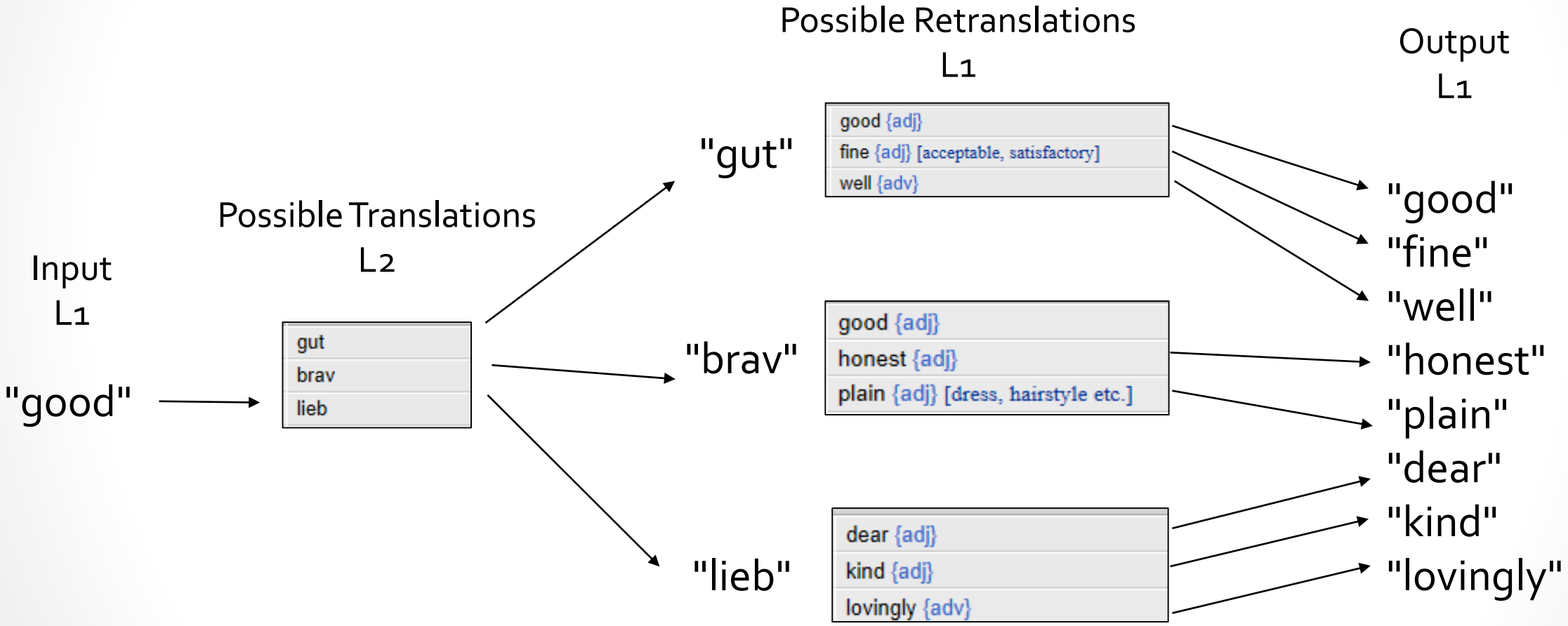
Clean Data
- Remove punctuation, etc.

Expand Data
- **Expand words**
- **Don't expand articles, etc.**

Calculate Models
- Word frequency

Number of Terms in Model by Topic			
Business	Politics	Science	Sports
3,873	4,217	5,509	9,049

Bilingual Dictionary Word Expansion (**RQ1**)



Large Corpus Training Data

Collect Data

- **Many documents per topic**

Clean Data

- Remove punctuation, etc.

Calculate Models

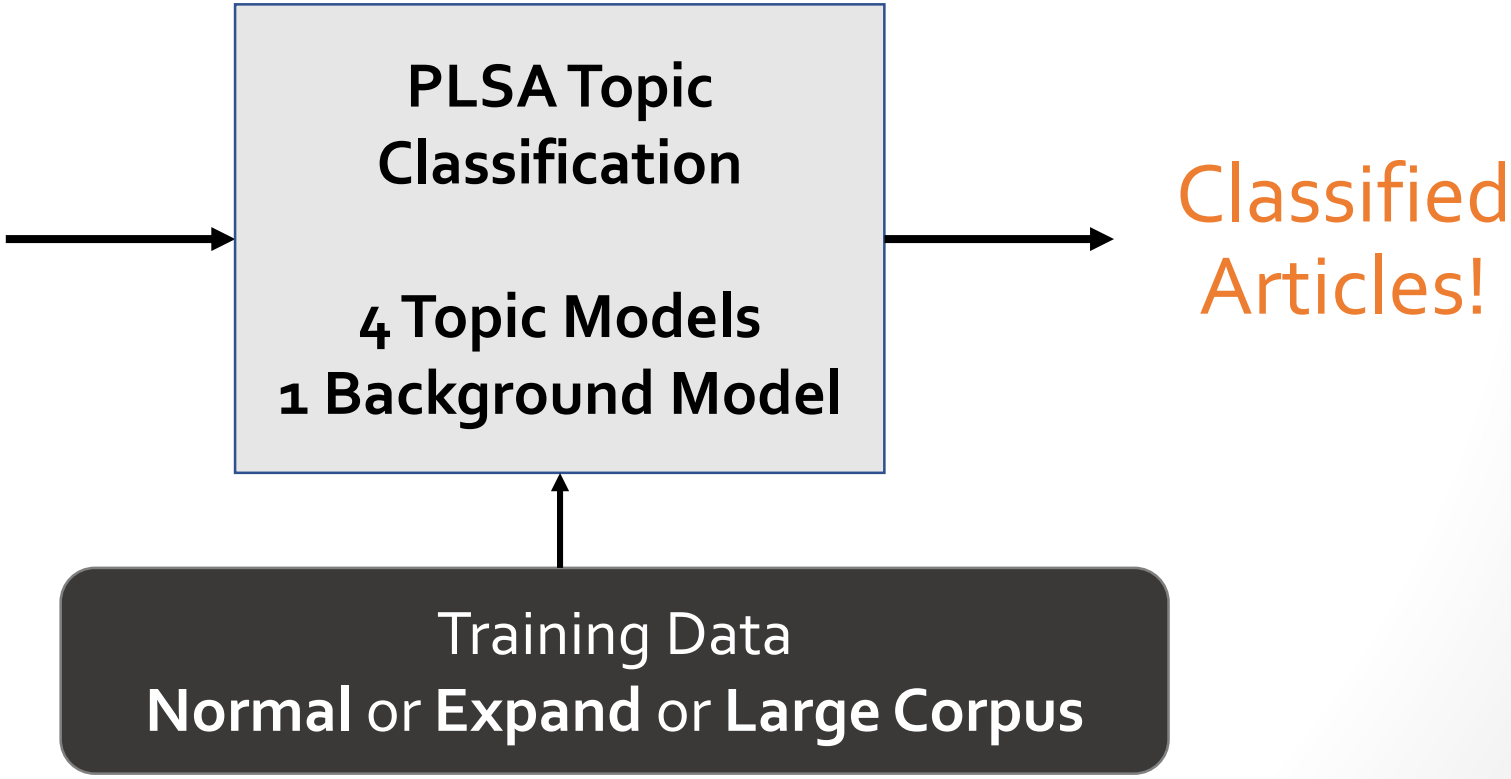
- Word frequency
- **TF-IDF**

Number of Terms in Model by Topic			
Business	Politics	Science	Sports
48,619	47,563	96,730	49,810

Testing

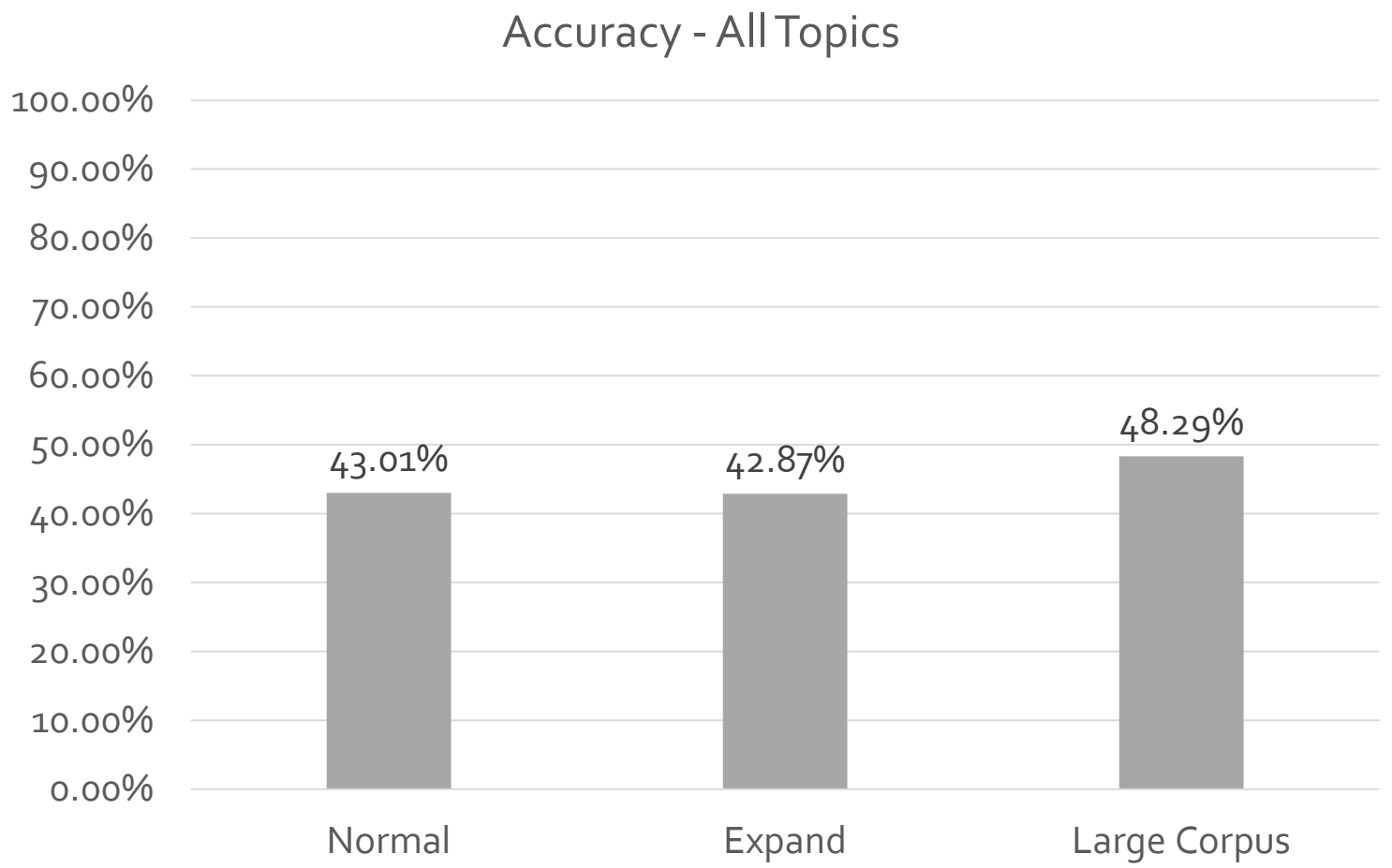
- News articles on 4 topics
  - Model classifies article from most probable topic model

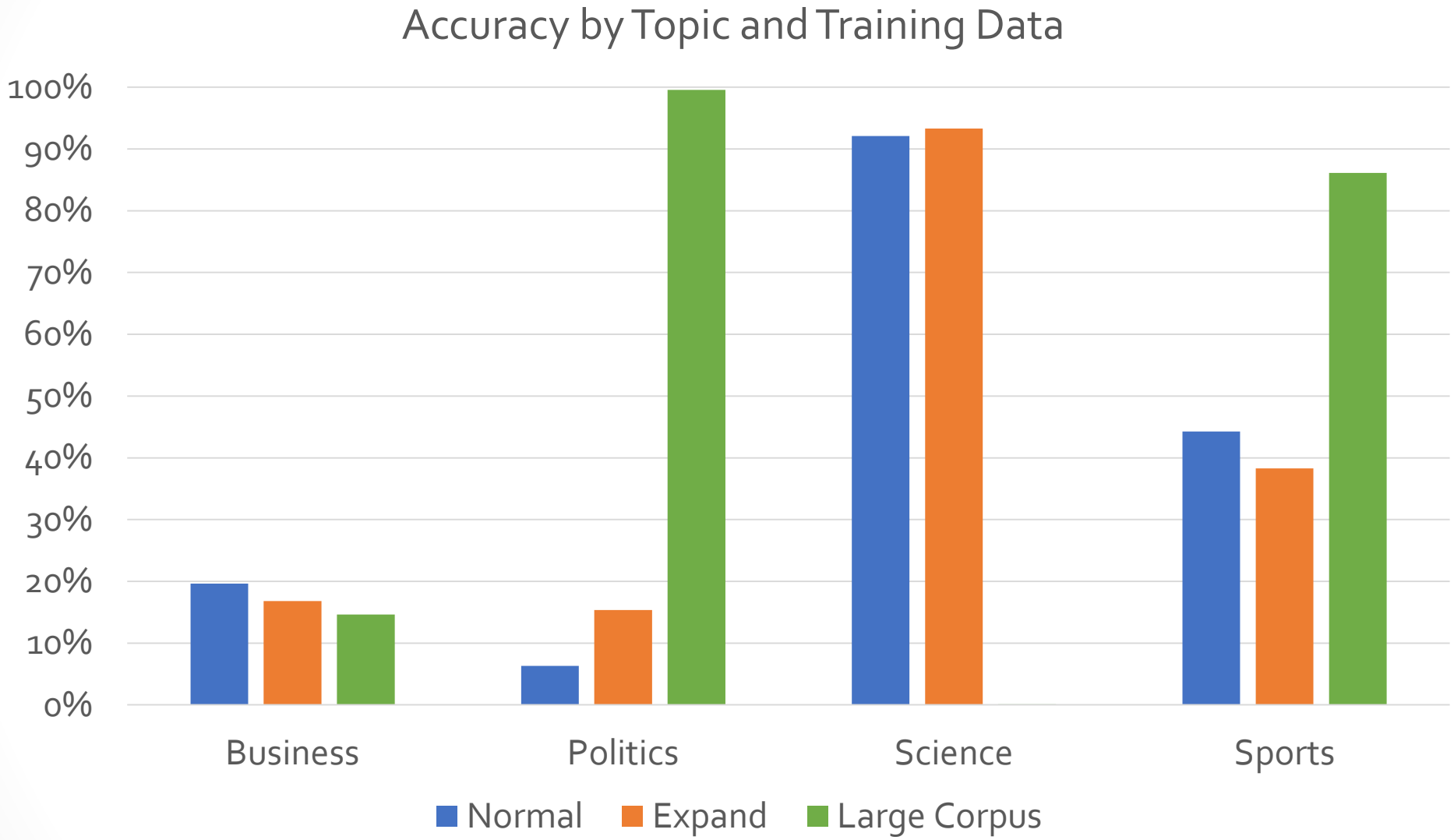
- Business (n = 780)
- Politics (n = 717)
- Science (n = 959)
- Sports (n = 1004)



Results

Accuracy = # Correct / # Total

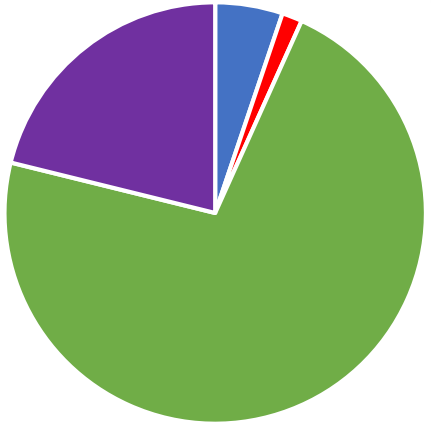



Accuracy by Topic and Training Data
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Business
Politics
Science
Sports
Normal
Expand
Large Corpus

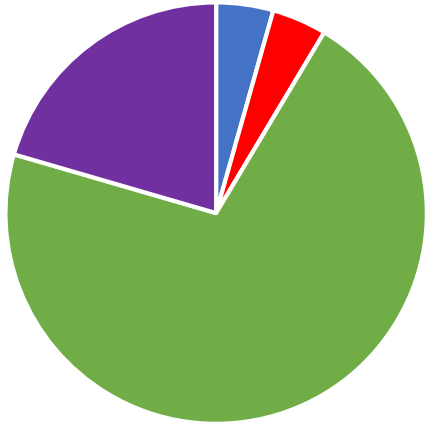
Distribution of Document Labels by Training Data Type

***Actual* # Labels**
Business (n = 780)
Politics (n = 717)
Science (n = 959)
Sports (n = 1004)

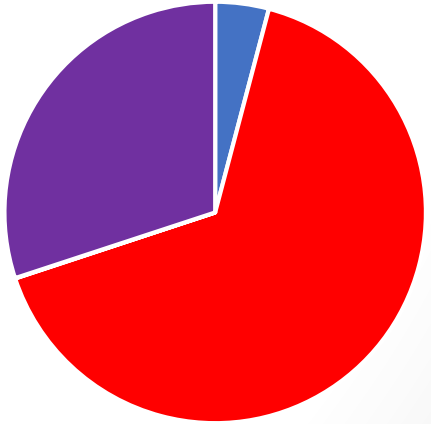
Normal # Labels
Business (n = 174)
Politics (n = 52)
Science (n = 2421)
Sports (n = 710)



Expand # Labels
Business (n = 147)
Politics (n = 141)
Science (n = 2380)
Sports (n = 688)

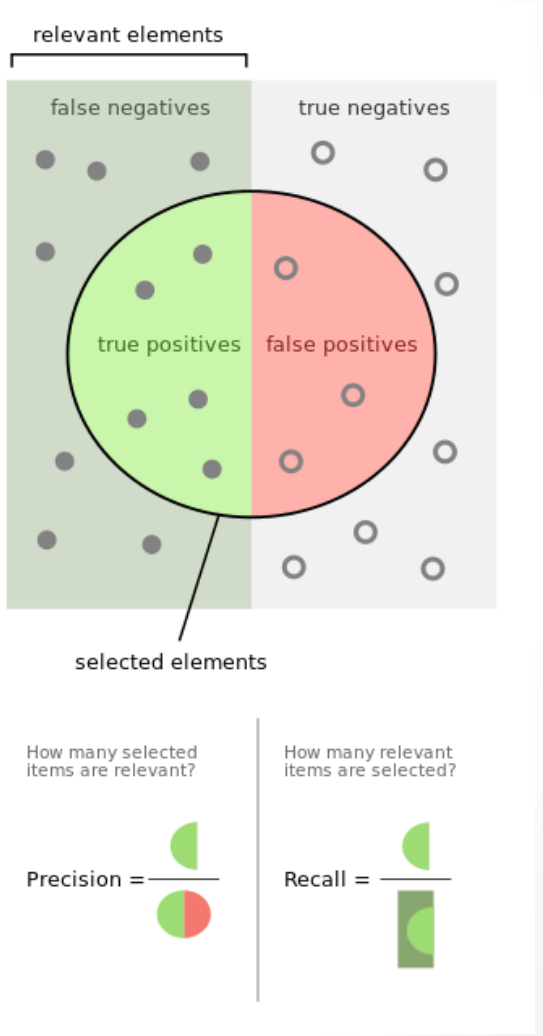


Large Corpus # Labels
Business (n = 138)
Politics (n = 2210)
Science (n = 1)
Sports (n = 1009)

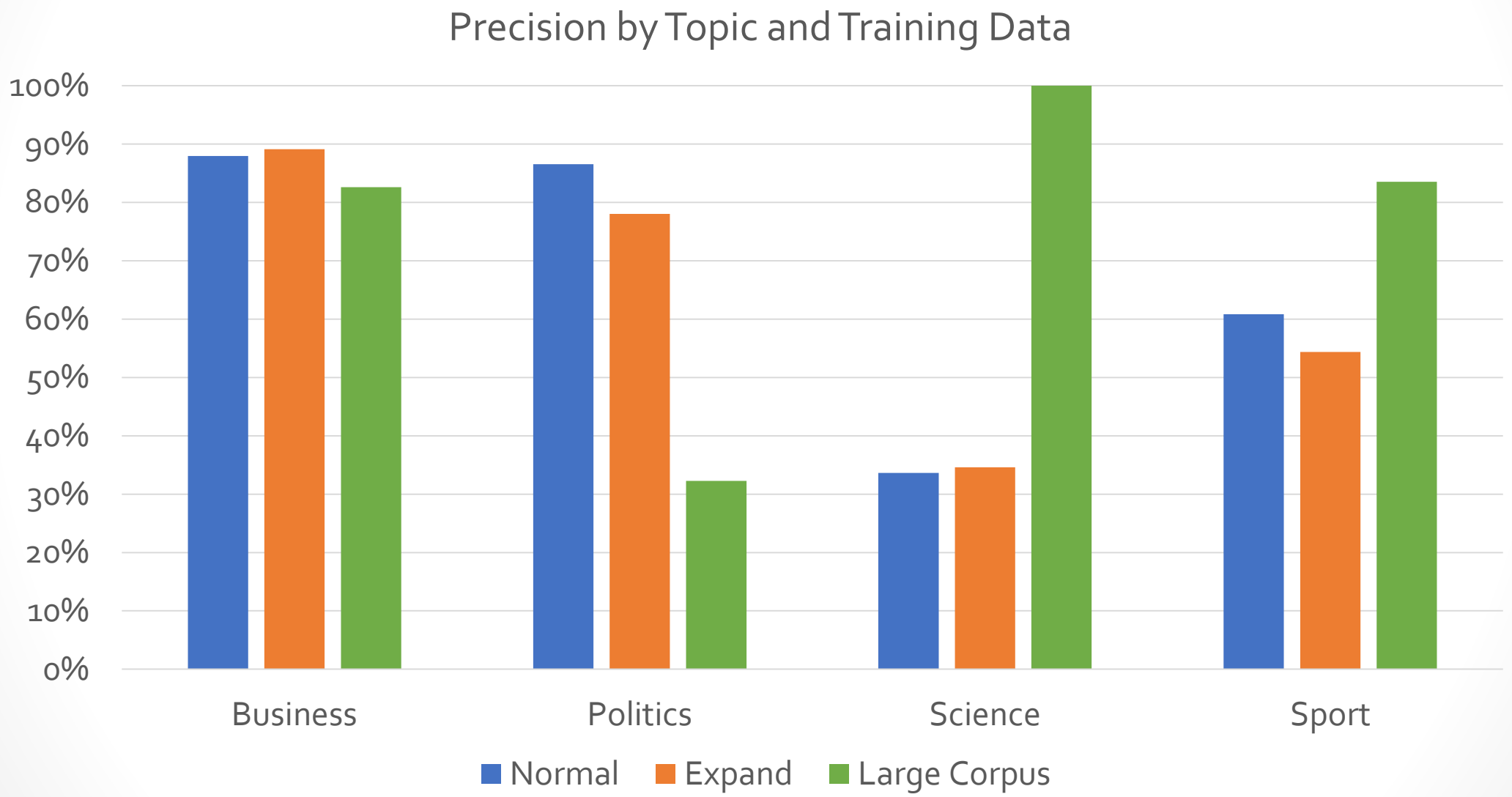


Precision, Recall, and F1 Scores [9]

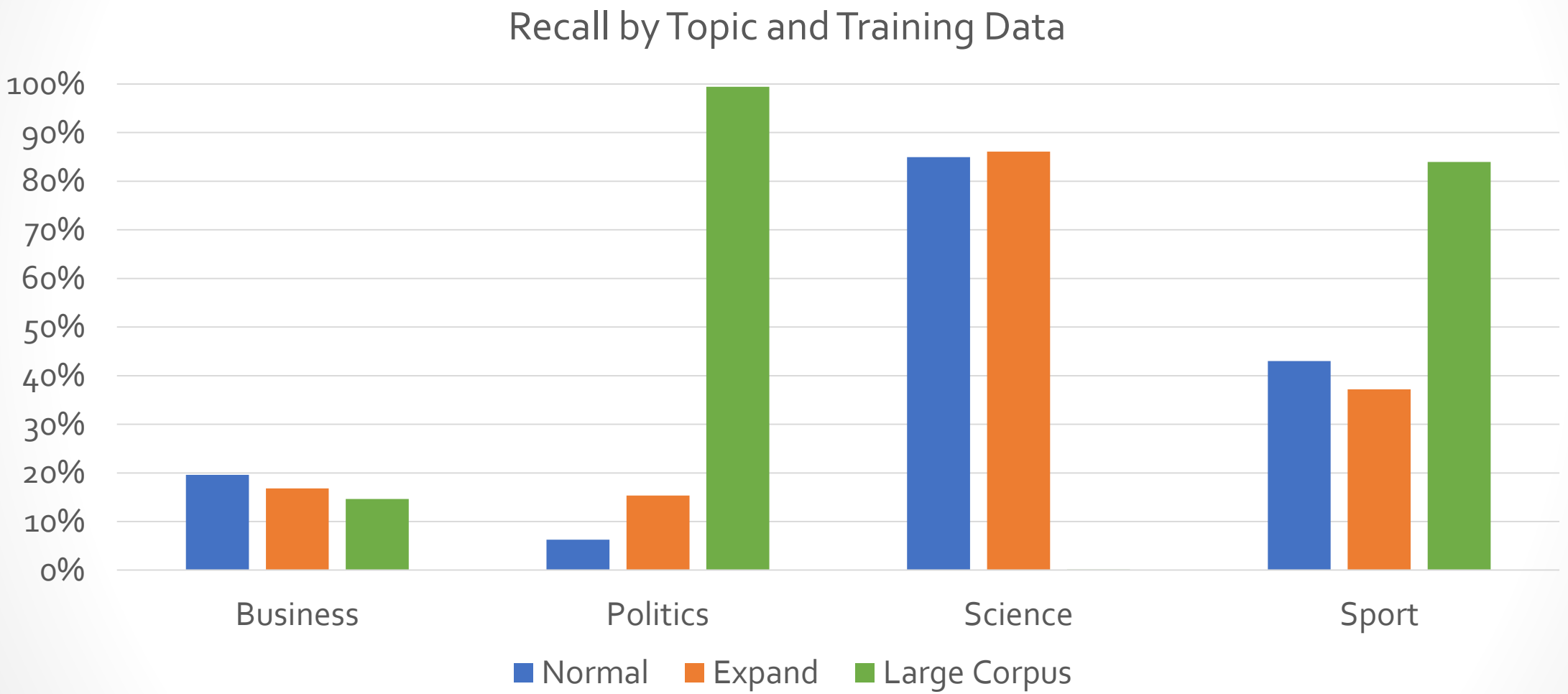
- Precision
  - What % of our **labeled data** is correct?
- Recall
  - What % of our **total data** is correctly labeled?
- F1 Score
  - Harmonic mean of Precision and Recall



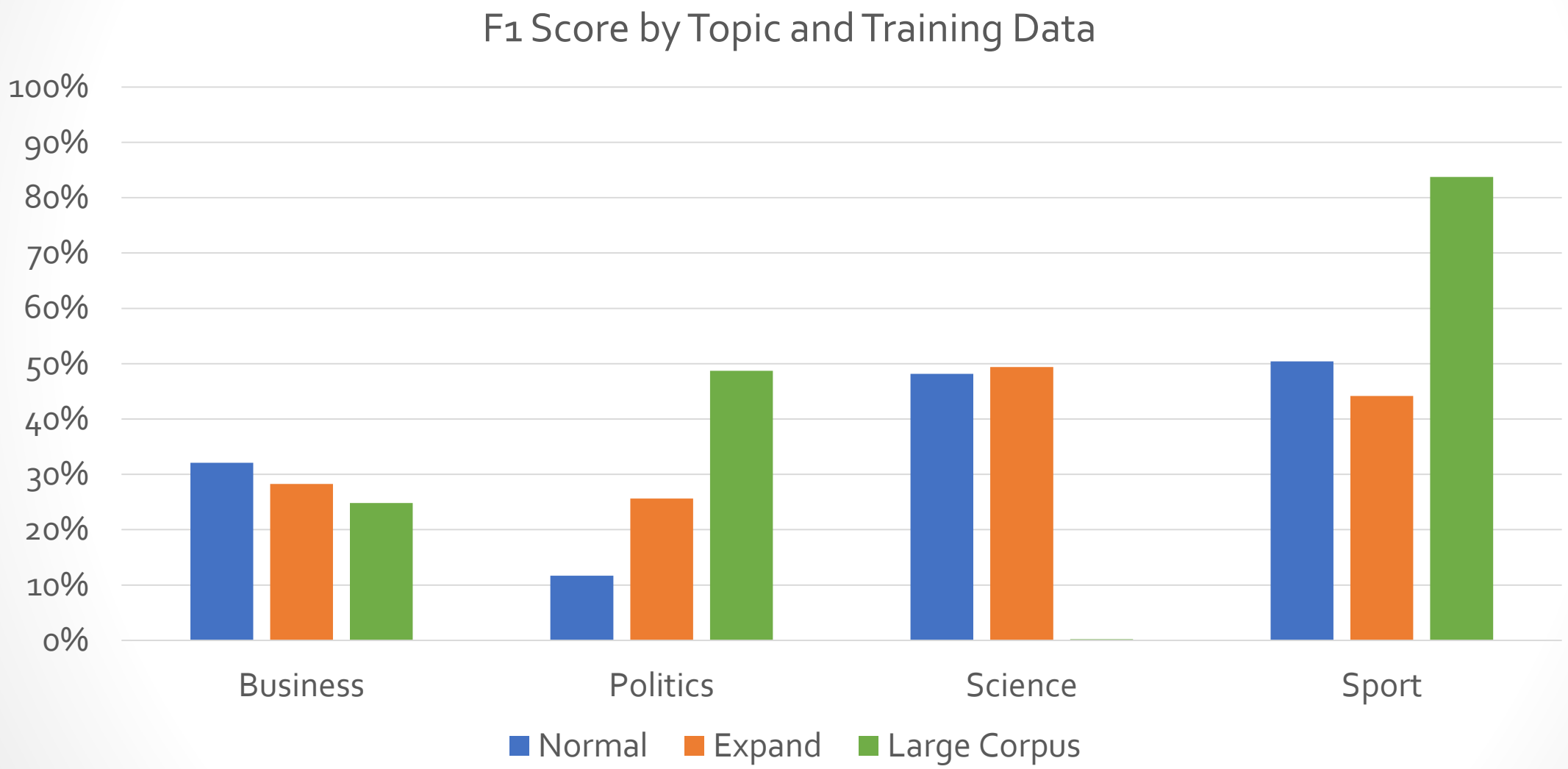
Precision = What % of our **labeled data** is correct?



Recall = What % of our **total data** is correctly labeled?



F1 Score = Harmonic Mean of Precision and Recall



Discussion

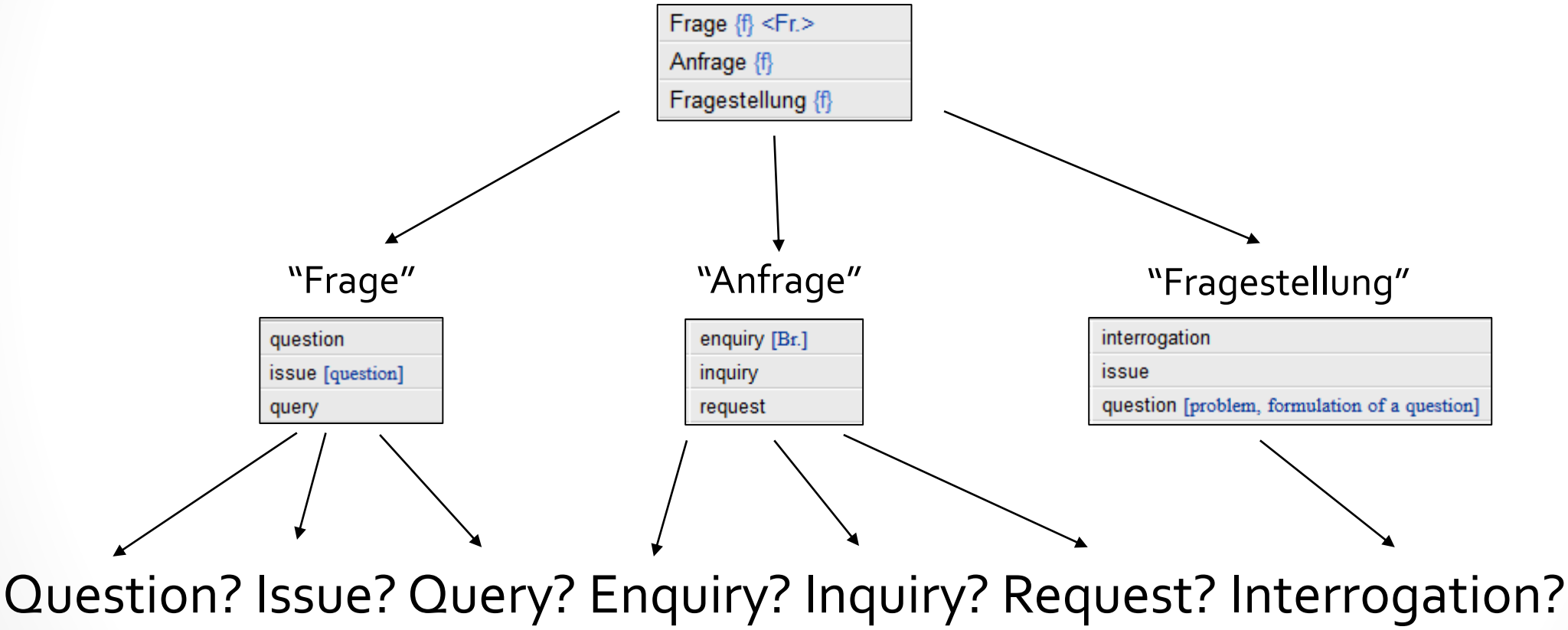
How did Expanded training data affect performance? (**RQ2**)

- Expanded performance was similar to Normal/Control performance
  - Similar misclassification problems
  - Inconsistently improved and hurt performance
    - Depended on the topic
    - Insignificant differences
- Expanded performance was NOT similar to Large Corpus performance
  - It is NOT a replacement for a large amount of data

Future Work

- Wider range of Topics
- Use standardizes experiment or dataset
- Larger expansion amount (only expanded with first 3 translations)
- Try with actual Low Resource Language
  - English-German was used for data accessibility
- Larger starting training data?
  - More than 8,00-2,300 words
- Further preprocessing before expansion

Questions?


Frage {f} <Fr.>
Anfrage {f}
Fragestellung {f}
“Frage”
“Anfrage”
“Fragestellung”
question
issue [question]
query
enquiry [Br.]
inquiry
request
interrogation
issue
question [problem, formulation of a question]
Question? Issue? Query? Enquiry? Inquiry? Request? Interrogation?

References

- [1] J. Ramapraba, S. Das, and P. Mukerjee, Survey on Sentence Similarity Evaluation using Deep Learning, Journal of Physics: Conference Series, vol. 1000, no. 1. IOP Publishing, 2018
- [2] C. Manning and R. Socher, Natural Language Processing with Deep Learning, Lecture at Stanford University, USA. 2017 [Online]. Available: https://www.youtube.com/watch?v=ASn7ExxLZws
- [3] S. Kim, K. Toutanova, and H. Yu. Multilingual Named Entity Recognition using Parallel Data and Metadata from Wikipedia, Association for Computational Linguistics, July, 2012.
- [4] A. Saif, N. Omar, M. J. Ab Aziz, U. Z. Zainodin and N. Salim, Semantic concept model using Wikipedia semantic features, Journal of Information Science, vol. 44, no. 4, pp. 526-551, 2018.
- [5] https://github.com/first20hours/google-10000-english
- [6] New York Times Article Search API. New York Times. https://developer.nytimes.com/article search v2.json#/README.
- [7] https://www.dict.cc
- [8] C. Zhai, Probabilistic Latent Semantic Analysis PLSA, Lecture at University of Illinois at Urbana-Champaign, USA. 2016 [Online]. Available: https://www.youtube.com/watch?v=vtadpVDr1hM
- [9] D. H. Kraft and E. Calvin, Fuzzy Information Retrieval, Morgan & Claypool Publishers, 2017.