11791: Milestone 1

Team 1:

Ran Chen, Pallavi Baljekar,

Jung In Lee, Chen Sun, Wei Zhang

Pipeline Structure

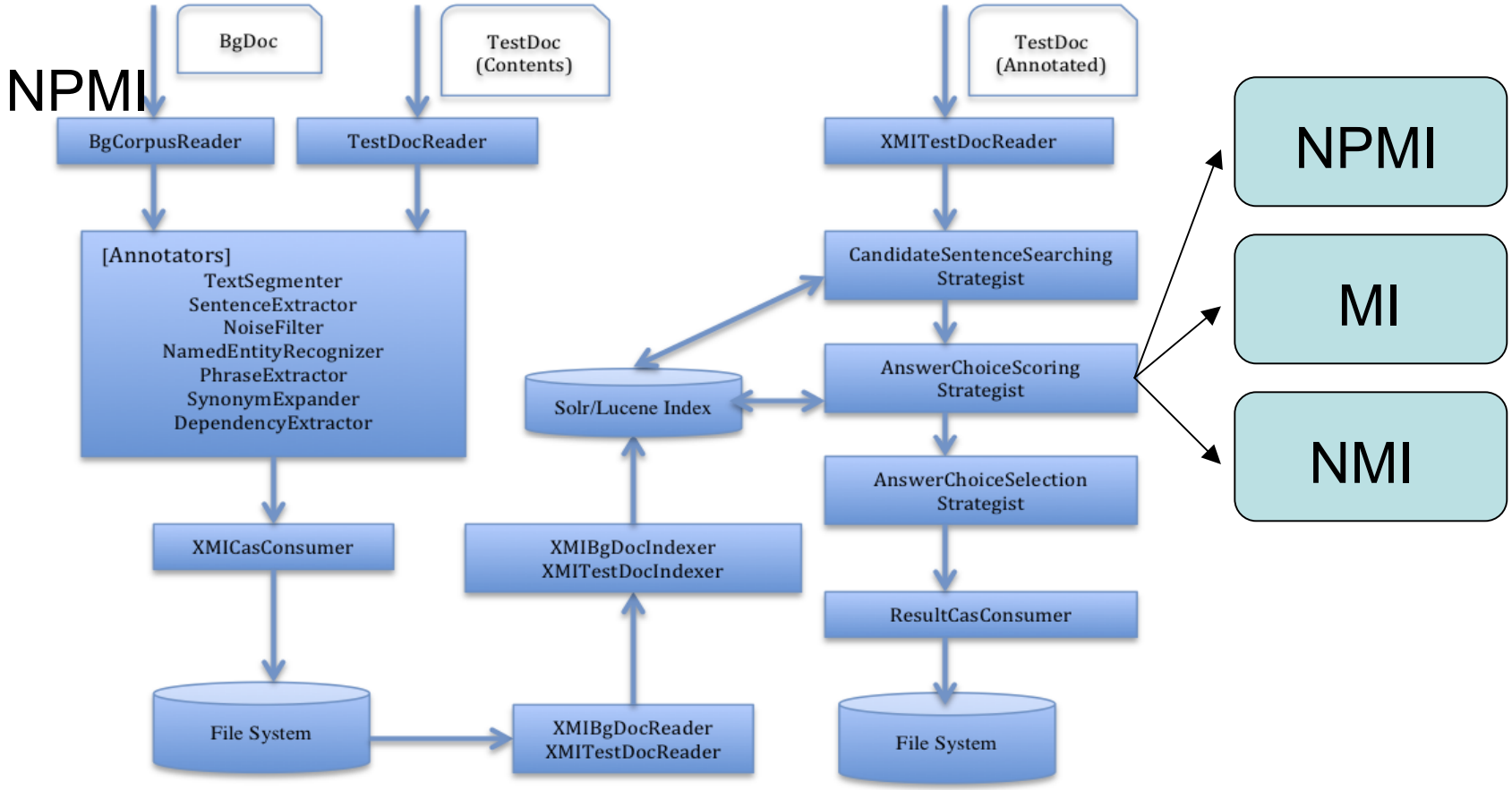

Figure 1. UIMA-based System Architecture for QA4MRE

Type System

- Maintaining the same Type System as given.

Baseline Methods

- PMI $$PMI(X,Y) = \ln \frac{p(x,y)}{p(x)p(y)}$$
  - Captures correlation between x, y.
  - Becomes infinitely large when x, y are rare.
- NPMI $$NPMI(X,Y) = \frac{\ln \frac{p(x,y)}{p(x)p(y)}}{-\ln p(x,y)}$$ (Gerlof 2009)
  - Reduce the impact of low frequency problem
  - Has nice lower & upper bounds:
    - When two words only occur together: 1
    - When they only occur separately: -1
    - When they are independent: 0

Baseline Methods

$$NPMI = \frac{\log\left(\frac{P(x,y)}{P(x)P(y)}\right)}{-\log\left(P(x,y)\right)}$$

$$NPMI = \frac{\log\big(P(x,y)\big) - \log\ P(x) - \log\ P(y)}{-\log P(x,y)}$$

$$NPMI = \frac{\log C(x,y) - Z - \log C(x) + Z - \log C(y) + Z}{-\log C(x,y) + Z}$$

where $Z = \log\ C(all\_doc)$

Baseline Methods

- MI

$$MI(X,Y) = \sum_{x,y} p(x,y) * PMI(x,y)$$

  - Expected value of PMI
  - Good indicator of whether two vectors are related or not (but not how much they're related)
- NMI

$$NMI(X,Y) = \frac{\sum_{x,y} p(x,y) * PMI(x,y)}{-\sum_{x,y} p(x,y) \ln p(x,y)}$$

  - Normalize by the joint entropy
  - Behaves more like PMI or NPMI
  - Recommended that MI and NMI be used together

(Gerlof 2009)

Performance

- Original Baseline: P(xly) 0 2 5 3 3 (**0.325)**
- PMI: ln(P(x,y)/(P(x)*P(y))) 0 4 2 4 1 (0.275)
- NPMI with Z: 0 4 1 2 2 (0.25)
  where Z = C(x)+C(y)-C(x,y)
- NPMI with Z^2: 0 2 0 1 1 (0.1)
- MI with Z: 0 4 4 2 3 (**0.325)**
- NMI with Z: 0 1 2 3 3 (0.225)
- P(x,y)^2/(P(x)*P(y)) 0 2 2 1 3 (0.1)
- 0.5 MI + 0.5 P(xly): 0 4 3 1 2 (0.25)

Future Plans

- Use Machine Learning methods to estimate the best coefficients
  - E.g. SVM
  - Challenge: Not enough data??
- Other knowledge-based approaches:
  - Synonyms, Acronyms…
  - Expanding NPs and/or other POS tags

Reference

- Gerlof, B. (2009) Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Chiarcos, Eckart de Castilho Stede (eds), Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, pp3140, Tbingen, Gunter Narr Verlag.
- Evert, S. (2004/2005) The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, IMS Stuttgart.