# Some papers in CICling2014

# 提纲

- Extracting Social Events based on Timeline and User Reliability Analysis on Twitter
- Bilingually Learning Word Senses for Translation
- Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge
- How Document Properties affect Document Relatedness Measures
- Credible or Incredible? Dissecting Urban Legends

## Extracting Social Events based on Timeline and User Reliability Analysis on Twitter （Chonbuk National University, Republic of Korea）

- To extract reliable low-frequency events as well as high-frequency events
- Propose an event extraction method based on timeline and user behavior analysis.

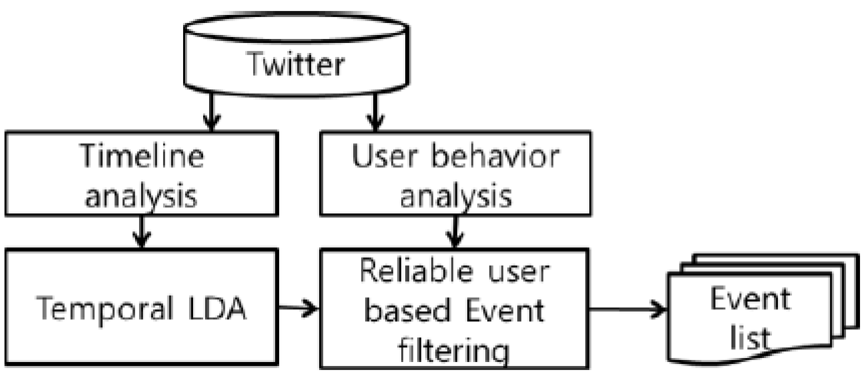



**Fig. 1.** The system structure of reliable user based event extraction

# Event extraction based on temporal LDA model

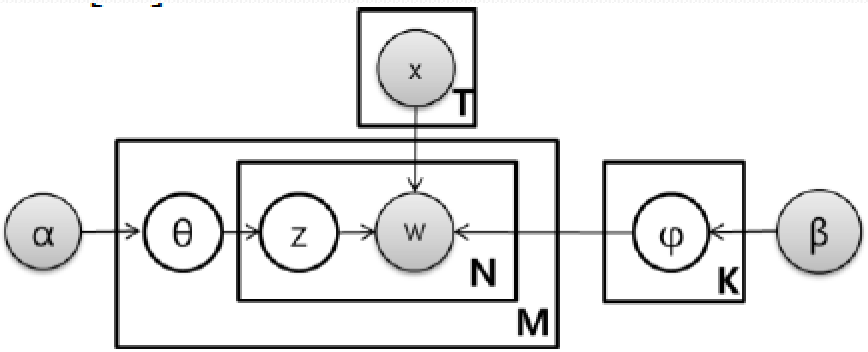



**Fig. 2.** Graphical representation of T-LDA model

## Reliable user detection

- Detecting socially well-known users.
  - tend to have a lot number of tweets and retweets.
  - HITS algorithm

$$AuthScore^{(T+1)}(p) = \sum_{q \to p} w_{qp} \times HubScore^{T}(q) \quad (2)$$

$$HubScore^{(T+1)}(p) = \sum_{p \to q} w_{pq} \times AuthScore^{T}(q) \quad (3)$$

The edge weight $w_{qp}$ is as follows:

$$w_{qp} = \sum_{q \to p} FreqRT(q,p) + \sum_{q \to p} Mention(q,p) \quad (4)$$

- Detecting active users.

$$Activity\ Score(u) = \frac{1}{W} \sum_{i=1}^{W} TweetFreq(u, d_i) \times RTFreq(u, d_i) \quad (5)$$

# Event filtering based on reliable users

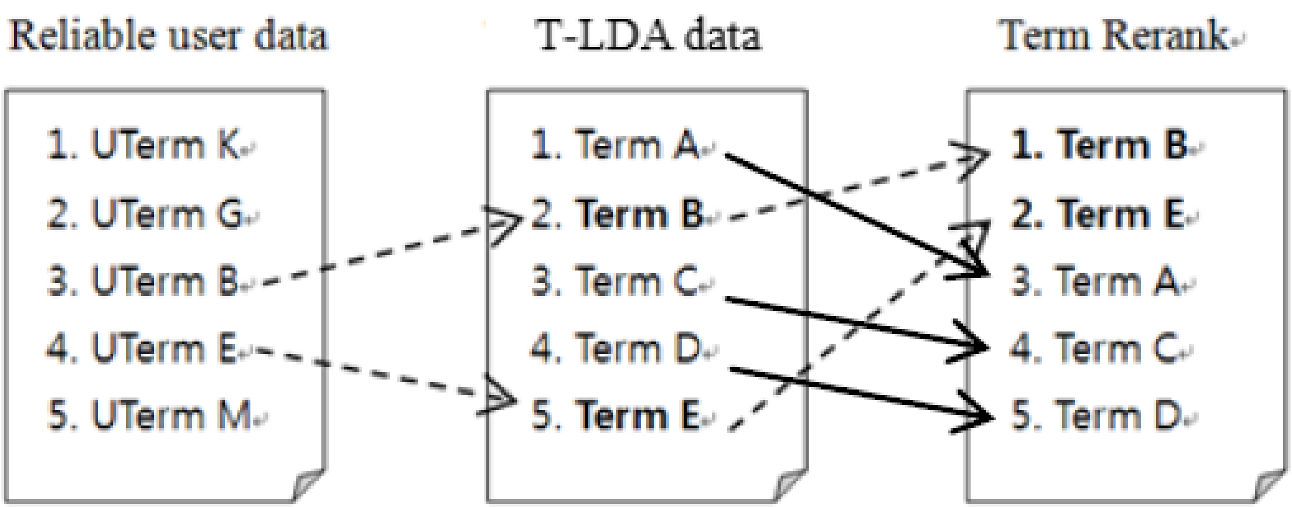



**Fig. 3.** Event filtering process

**Table 6.** Summary of comparison results (P@10)

| *Issue No* | Chi-OpScore | T-LDA | Proposed method | *Issue No* | Chi-OpScore | T-LDA | Proposed method |
|---|---|---|---|---|---|---|---|
| E1 | **5/5** | **5/5** | **5/5** | E13 | **3/3** | **3/3** | **3/3** |
| E2 | **4/4** | **4/4** | **4/4** | E14 | **2/2** | **2/2** | **2/2** |
| E3 | **4/4** | **4/4** | **4/4** | E15 | **4/4** | **4/4** | **4/4** |
| E4 | **5/5** | **5/5** | **5/5** | E16 | 5/6 | 5/6 | **6/6** |
| E5 | 6/7 | 6/7 | **7/7** | E17 | **3/3** | **3/3** | **3/3** |
| E6 | **2/2** | **2/2** | **2/2** | E18 | **1/1** | **1/1** | **1/1** |
| E7 | **4/4** | **4/4** | **4/4** | E19 | **3/3** | **3/3** | **3/3** |
| E8 | **3/3** | **3/3** | **3/3** | E20 | **2/2** | **2/2** | **2/2** |
| E9 | **5/6** | **5/6** | **6/6** | E21 | **2/3** | **2/3** | **2/3** |
| E10 | **3/3** | **3/3** | **3/3** | E22 | **5/5** | **5/5** | **5/5** |
| E11 | **2/3** | **2/3** | **2/3** | E23 | **6/6** | **6/6** | **6/6** |
| E12 | **3/3** | **3/3** | **3/3** | E24 | **4/4** | **4/4** | **4/4** |
| | | | | Avg | 94.3% | 95.2% | 97.2% |

## Bilingually Learning Word Senses for Translation

- learns word sense clusters and then uses learned contextual information for classifying expressions according to the sense of ambiguous words occurring there.
- Approach
  - Selection of Word Senses
    - ISTRION EN-PT lexicon
      - 850.000 English-Portuguese
  - Features Extraction
    - Parallel corpus, window
  - Features Correlation
  - Clusters Construction
    - X-means

# Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge

- Present a bilingual lexicon extraction system that is based on a novel combination of topic model and context based methods.

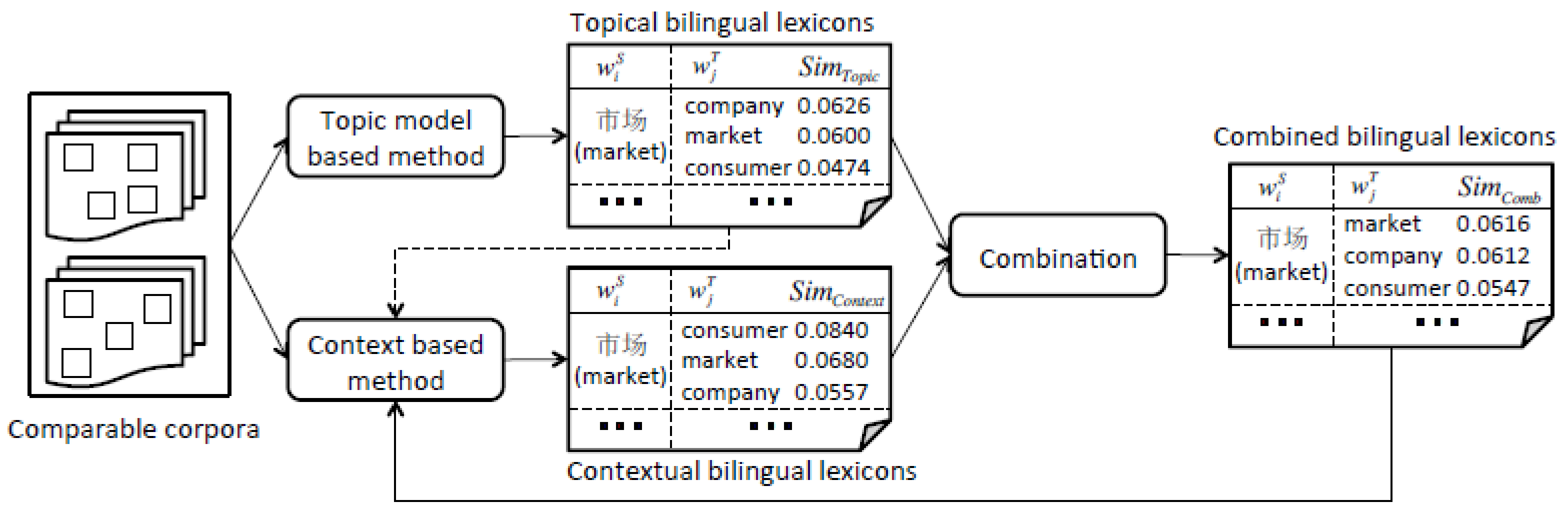



Fig. 1. Bilingual lexicon extraction system.

# Topic Model Based Method

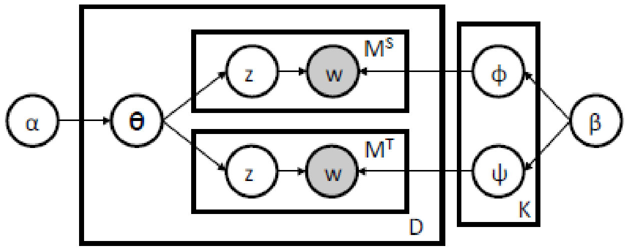



**Fig. 2.** The BiLDA topic model.

- *TI+Cue* measure

$$Sim_{TI+Cue}(w_i^S, w_j^T) = \lambda Sim_{TI}(w_i^S, w_j^T) + (1-\lambda)Sim_{Cue}(w_i^S, w_j^T)$$

- *TI* measure
  - Source and target word vectors constructed over a shared space of cross-lingual topics.
  - Each dimension of the vectors is a *TF-ITF* (term frequency -inverse topic frequency) score.
  - Cosine similarity
- *Cue* measure

$$P(w_j^T | w_i^S) = \sum_{k=1}^{K} \psi_{k,j} \frac{\phi_{k,i}}{Norm_\phi} \tag{4}$$

where $Norm_\phi$ denotes the normalization factor given by $Norm_\phi = \sum_{k=1}^{K} \phi_{k,i}$ for a word $w_i$.

## Context Based Method

- Window-based context
  - window size of 4
  - TF-IDF
  - project the source vector onto the vector space of the target language using a seed dictionary.
  - Cosine similarity

# Combination

$$Sim_{Comb}(w_i^S, w_j^T) = \gamma Sim_{Topic}(w_i^S, w_j^T) + (1-\gamma) Sim_{Context}(w_i^S, w_j^T)$$

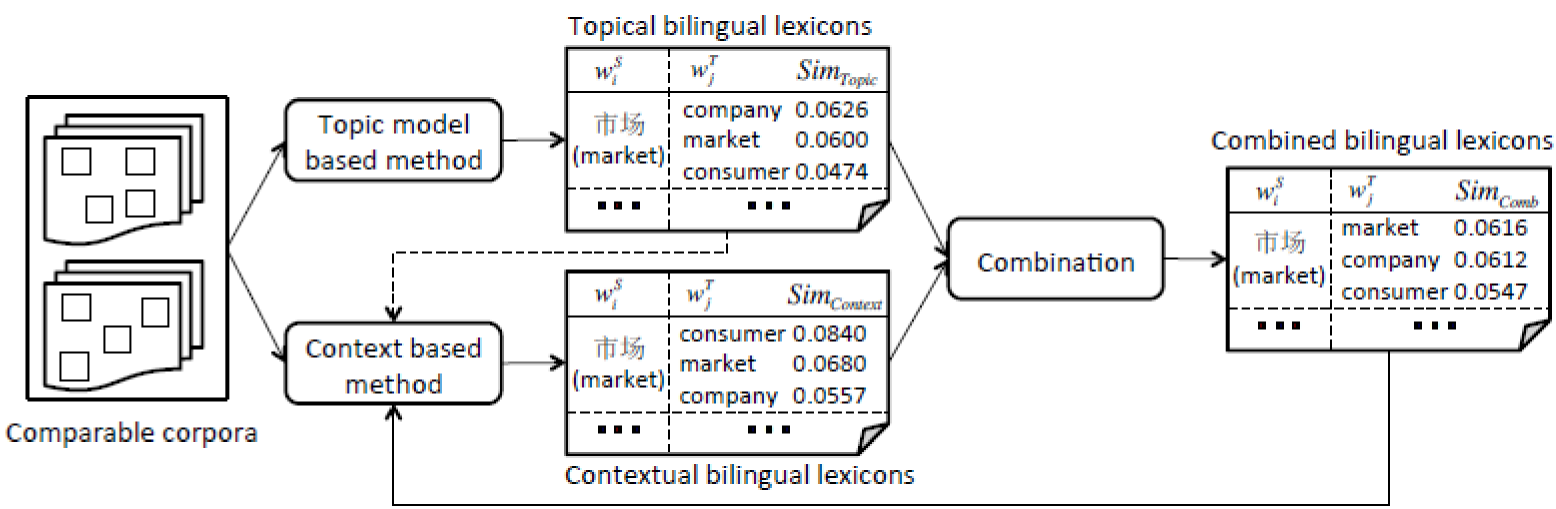



Fig. 1. Bilingual lexicon extraction system.

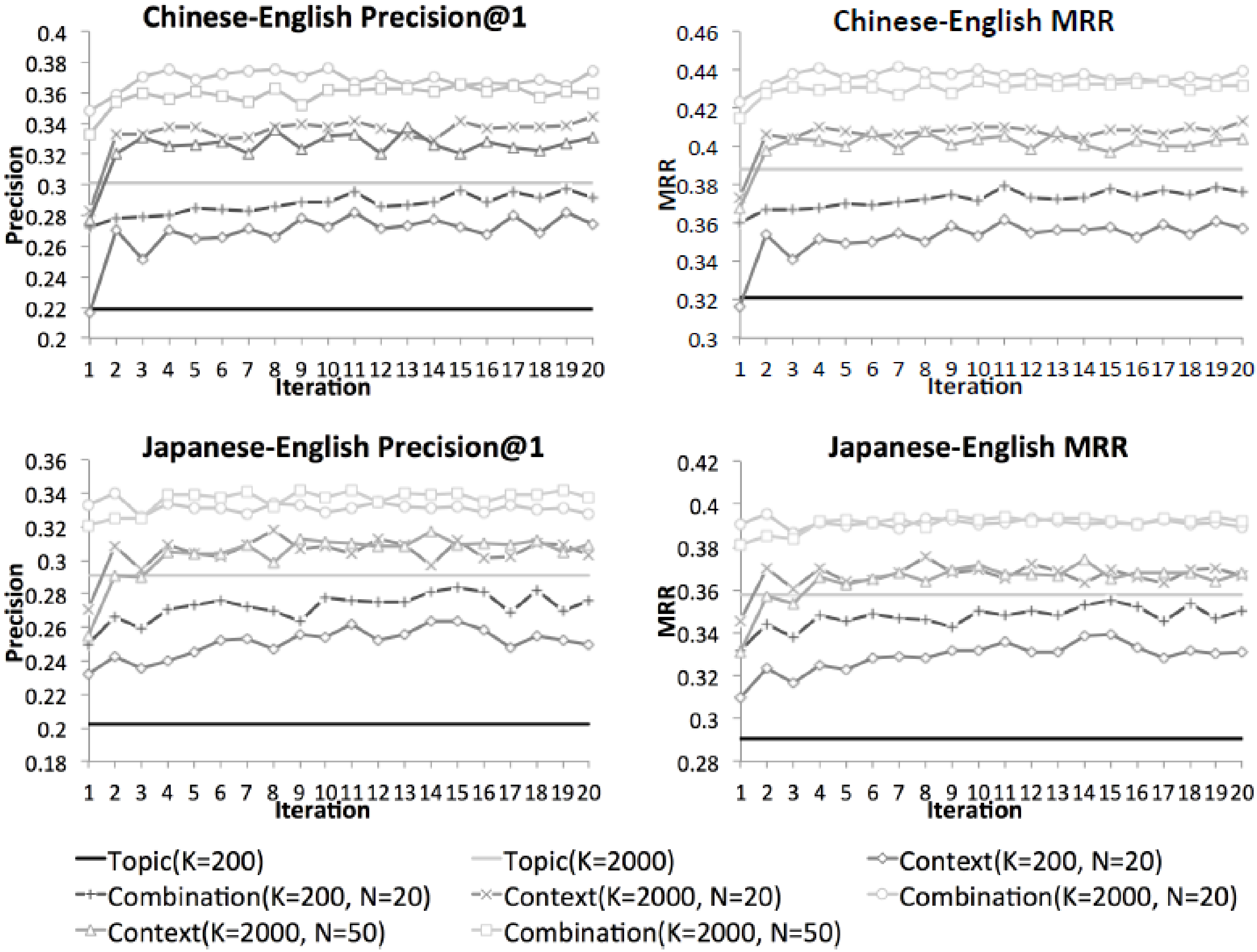
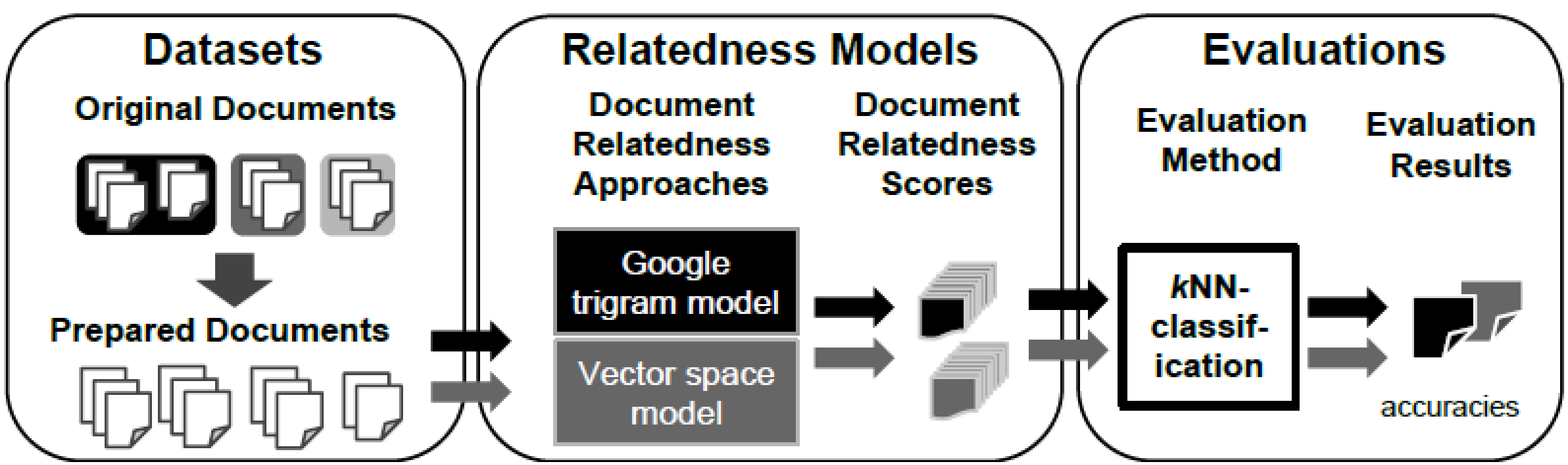



**Fig. 3.** Results for Chinese–English and Japanese–English on the test sets.

*K* denotes the number of topics and *N* denotes the number of translation candidates for a word we compared in our experiments.

# How Document Properties affect Document Relatedness Measures(1)

- how document properties (word count, term frequency, cohesiveness, genre) affect the quality of unsupervised document relatedness measures (Google trigram model and vector space model).



# How Document Properties affect Document Relatedness Measures(2)

- Dataset
  - Aviation Safety Reporting System (ASRS).
    - 399 ASRS reports, 96 words on average
    - Incursion (collision hazard) (165), Altitude deviation (59), Fire or smoke problems (62), and Security Concern Threat (116)
  - Medical Vigilance Report List (Med)
    - 659 vigilance reports, 19 words on average
    - Software (298) or hardware (361)
  - Biodiversity Heritage Library (BHL).
    - Titles
      - 1152 titles, 7 words on average
      - Poultry (297), Zoology (289), Agriculture (297),Botany (269)
    - Introductions
      - 338, 152 words on average
      - Sheep (58), Biochemistry (63), Dairying(64), Bacteriology (94), Tobacco (59).

# How Document Properties affect Document Relatedness Measures(3)

- Document Relatedness Models
  - Vector Space Model (VSM).
  - Google Trigram Model (GTM).

$$\frac{(\delta + \sum_{i=1}^{|d_1|-\delta} \mu(A_i)) \times (|d_1| + |d_2|)}{2|d_1||d_2|} \qquad (1)$$

- kNN-Classication
- Document Attribute Values
  - Word Count: The number of words within a document.
  - Term Frequency: A normalized average of the frequency of each word
  - Cohesion: The average word similarity

# How Document Properties affect Document Relatedness Measures(4)

**Table 1.** $k$NN-classification 10-fold cross-validation result summary for each attribute at limits in the minimum lower bound (Min.), maximum upper bound (Max.), interval (Int.). The percentage of tests in which 1-sided significance is found, is shown under "GTM ? VSM". The correlation coefficients between the average attribute values of each dataset subset and the mean classification accuracy are presented (Attr. Correlation) following different relation patterns: Positive linear (Pl), Negative linear (Nl), Positive parabolic (Pp), and Negative parabolic (Np). Highest correlations of each approach are **bolded**.

| | Limits | | | GTM ? VSM | | | Attr. Correlation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Min. | Max. | Int. | > | < | no diff. | GTM | | VSM | |
| *Word Count:* | | | | | | | | | | |
| ASRS | 6 | 302 | 8 | 36.6 | 41.7 | 21.7 | Pl | **0.662** | Np | 0.366 |
| Med | 2 | 100 | 2 | 62.2 | 26.0 | 11.8 | Pp | 0.531 | Pp | **0.603** |
| BHL Titles | 0 | 36 | 2 | 67.5 | 14.2 | 18.3 | Pp | 0.004 | Pp | **0.031** |
| BHL Intro | 53 | 539 | 9 | 0.0 | 99.0 | 0.1 | Nl | 0.335 | Nl | **0.625** |
| *Term Frequency:* | | | | | | | | | | |
| ASRS | 0.04 | 0.36 | 0.01 | 17.5 | 57.3 | 25.2 | Np | **0.713** | Np | 0.561 |
| Med | 0.01 | 0.52 | 0.01 | 68.0 | 23.6 | 8.4 | Pl | 0.721 | Pl | **0.931** |
| BHL Titles | 0.00 | 1.00 | 0.05 | 63.8 | 30.7 | 5.5 | Np | **0.604** | Np | 0.578 |
| BHL Intro | 0.03 | 0.21 | 0.01 | 1.0 | 91.0 | 8.0 | Pp | **0.859** | Pp | 0.834 |
| *Cohesion:* | | | | | | | | | | |
| ASRS | 0.15 | 0.30 | 0.01 | 20.8 | 65.3 | 13.9 | Np | **0.889** | Np | 0.882 |
| Med | 0.00 | 0.37 | 0.01 | 74.1 | 17.3 | 8.6 | Np | 0.276 | Np | **0.620** |
| BHL Titles | 0.00 | 0.45 | 0.01 | 79.5 | 9.3 | 11.2 | Np | **0.517** | Np | 0.470 |
| BHL Intro | 0.05 | 0.35 | 0.01 | 0.0 | 99.3 | 0.0 | Np | **0.743** | Np | 0.719 |

## Credible or Incredible?
## Dissecting Urban Legends（1）

- Urban legends are a genre of modern folklore, consisting of stories about rare and exceptional events, just plausible enough to be believed.

Table 1. Examples of Urban Legend Claims

| |
|---|
| A tooth left in a glass of Coca-Cola will dissolve overnight. |
| A stranger who stopped to change a tire on a disabled limo was rewarded for his efforts when the vehicle's passenger, Donald Trump, paid off his mortgage. |
| Walt Disney arranged to have himself frozen in a cryonic chamber full of liquid nitrogen upon his death, and he now awaits the day when medical technology makes his re-animation possible. |
| Drugged travelers awaken in ice-filled bathtubs only to discover one of their kidneys has been harvested by organ thieves. |
| Facebook users can receive a $5,000 cash reward from Bill Gates for clicking a share link. |

## Credible or Incredible? Dissecting Urban Legends （2）

- UL should mimic the details of news (who, where, when) to be credible, and they should be emotional and readable like the story of a fairy tale to be catchy and memorable.
- Dataset
  - Urban Legends, 5000
  - News Articles, 400.000 Google News articles
  - Fairy Tales, 1860
- Feature
  - NE, Temporal Expressions, Sentiment (SENT), Readability

# Credible or Incredible? Dissecting Urban Legends （3）

**Table 6.** Classification Results

| | UL vs. GN | | | UL vs. FT | | | GN vs. FT | | |
|---|---|---|---|---|---|---|---|---|---|
| Features | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 |
| NE | 0.694 | 0.694 | 0.694 | 0.787 | 0.768 | 0.777 | 0.897 | 0.896 | 0.896 |
| TIMEX | 0.677 | 0.676 | 0.676 | 0.666 | 0.666 | 0.666 | 0.775 | 0.767 | 0.766 |
| SENT | 0.573 | 0.572 | 0.572 | 0.661 | 0.656 | 0.658 | 0.606 | 0.601 | 0.603 |
| READ | 0.765 | 0.762 | 0.763 | 0.869 | 0.868 | 0.868 | 0.973 | 0.973 | 0.973 |
| ALL | 0.834 | 0.833 | 0.833 | 0.897 | 0.897 | 0.897 | 0.978 | 0.978 | 0.978 |

**Table 7.** Results for UL vs FT vs GN

| Features | Prec | Rec | F1 | MCC |
|---|---|---|---|---|
| NE | 0.630 | 0.650 | 0.640 | 0.449 |
| TIMEX | 0.570 | 0.577 | 0.573 | 0.339 |
| SENT | 0.446 | 0.461 | 0.453 | 0.069 |
| READ | 0.746 | 0.754 | 0.750 | 0.611 |
| ALL | 0.820 | 0.822 | 0.821 | 0.721 |
| ZeroR | 0.202 | 0.450 | 0.279 | 0 |

**Table 8.** Overall Feature performances

| Features | F1$\mu$ | F1$\sigma$ |
|---|---|---|
| ALL | 0.868 | 0.070 |
| READ | 0.819 | 0.100 |
| NE | 0.740 | 0.100 |
| TIMEX | 0.675 | 0.069 |
| SENT | 0.589 | 0.085 |

Thank you ！

Q&A