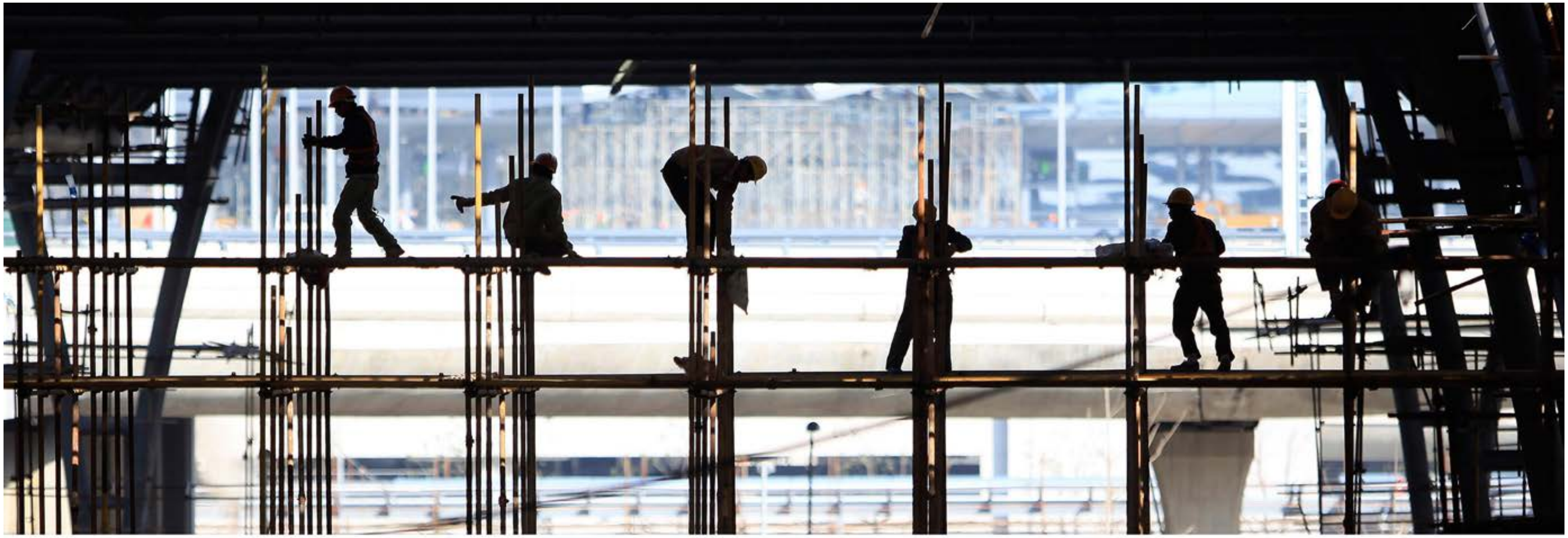


REUTERS/Aly Song

A Statistical NLG Framework for Aggregated Planning and Realization

Ravi Kondadadi, Blake Howald and Frank Schilder

Research & Development, Thomson Reuters

ACL 2013

Sofia, Bulgaria

August 6, 2013

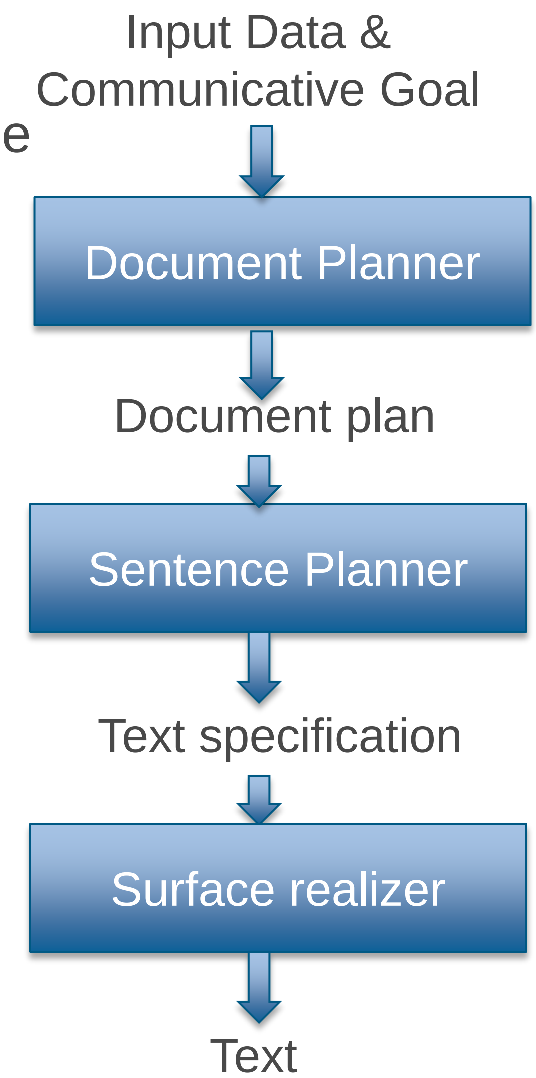
THOMSON REUTERS

OVERVIEW

- Background on approaches to NLG
- A statistical framework to NLG
  - Clustering templates
  - Learning sentence and document planning
- Developed systems applied to *Biography* and *Weather* domains
- Evaluations:
  - Automatic Metrics (BLEU-4, METEOR, *Syntactic Variability*)
  - Non-Expert crowdsourcing (CrowdFlower)
  - Expert evaluations (*Biography*)
- Conclusions

NLG SYSTEM ARCHITECTURE (Reiter & Dale 2000)

- Data/Communicative Goal
  - Provide textual information about some subject matter/ domain
- Document (*Macro*-) Planning
  - "What to say" (Content)
    - Content selection
    - Document structuring
- Sentence (*Micro*-) Planning
  - "How to say" (Sentence)
    - Word choice, phrase composition, pronoun use and resolution, etc.
- Surface realization
  - Putting everything together into "natural" sounding texts



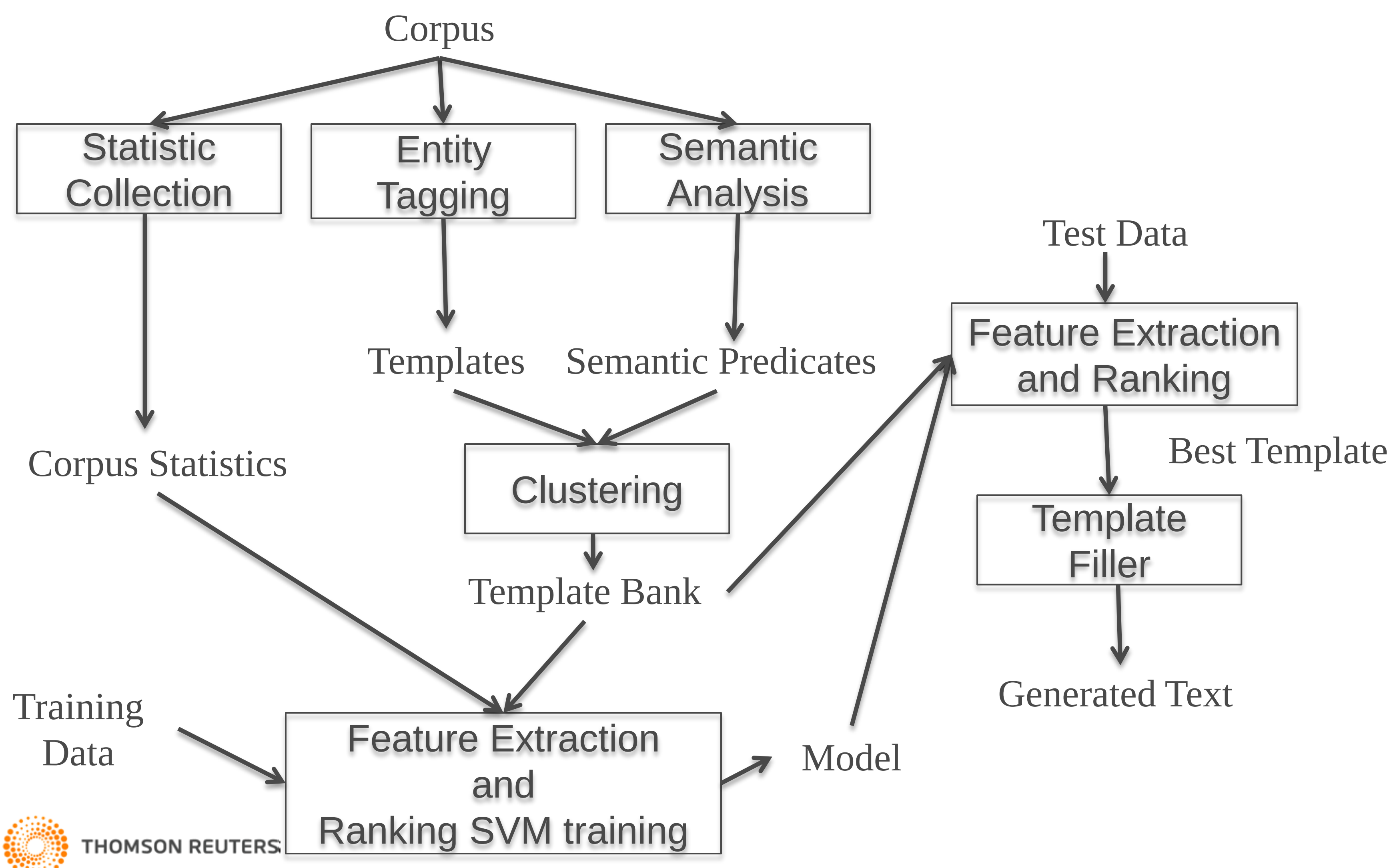
BACKGROUND

- Our system is a hybrid statistical-template system
  - Templates avoid necessity of an extensive grammar
  - Statistical approach, provided a robust corpus, allows for expedited learning of :
    - Content Selection - organization of the semantic structure of historical data
    - Document Planning – sequence of templates and domain tags
    - Sentence Planning – domain general and specific tagging and template generation
    - Surface Realization – template selection and content filling

ACL 2013

IWCS 2013

NLG FRAMEWORK



THOMSON REUTERS

CREATING TEMPLATE BANKS

- Automatic clustering (k-Means) and manual review creates template banks
- Template banks contains clusters of templates derived from corpus via NE tagging and semantic analysis:
  a. …
  b. ***[person]*** *holds a [degree] in [subject] from [school] and a [degree] from [school]*
  c. ***[person]*** *graduated from [school] with a degree in [subject]*
  d. ***[person]*** *graduated from [school] with a degree in [subject] and also a [degree] in [subject]*
  e. ***[person]*** *received a [degree] from [school] in* ***[date]***
  f. …
- Conceptual Units are manually assigned to clusters:
  - *CuId: 001 – "current position"*;
  - *CuId: 002 – "previous position"*;
  - *CuId: 003 – "education";*

THOMSON REUTERS

COLLECTION OF CORPUS STATISTICS

- Frequency distribution of templates overall and per position
- Frequency distribution of CuIds overall and per position
- Average number of words per CuId, position and combination of both
- Average number and distribution of entity tags by CuId
- Average number and distribution of entity tags by position
- Frequency distribution of CuId sequences (bigrams and trigrams)
- Frequency distribution of template sequences (bigrams and trigrams)
- Frequency distribution of entity tag sequences overall and per position
- The average, minimum, maximum number of CuIds across all documents

FEATURES FOR RANKING SVM

- Feature values are binary (1|0) or real values [0..1]
  - CuId given position
  - Overlap of named entities
  - Prior template/ CuId
  - Difference in number of words given position
  - Percentage of unused data/ Average number of words used
  - Difference in number of named entities
  - Average number of entities
  - Most likely CuId given position and previous CuId
  - Similarity between the most likely template in CuId and current template

RANKING

- Training (70%) – for each template in all training documents:
  1. All other templates in CuId (filtered by entities) are ranked by Levenshtein edit distance
  2. Corpus statistics used to calculate all features for each ranked template
  3. Ranking SVM assigns model weights to all features
- Testing / Generation (30%)
  1. Select most likely CuId for position 1 given input data
  2. Filter templates by input data
  3. Score all remaining templates (multiplying feature values by model weights)
  4. Select top scored template, fill input data
  5. Remove or modify used input data
  6. Repeat until input data exhausted (within average min/max length)

DATA

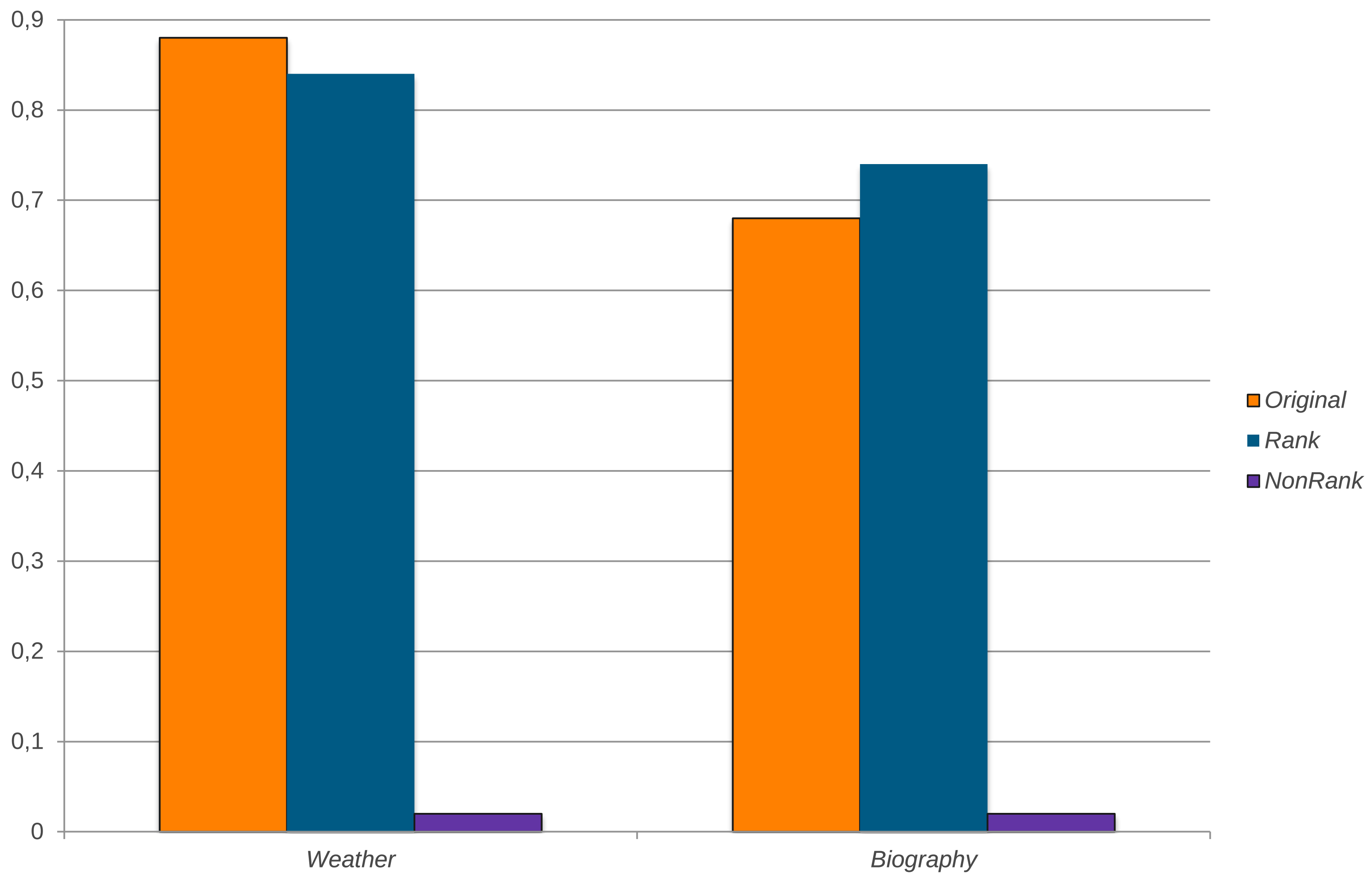
- *Biography* – Human generated (journalists)
  - Corporate officers and directors biographies
- *Weather* – Human generated (weather forecasters)
  - Offshore oil rig weather forecasts (SumTime-Meteo (Reiter et al. 2005))

	Biography	***Weather***
Texts (Sentence Range)	*1150 (3-17)*	*1045 (1-6)*
Conceptual Units	*19*	*9*
Templates	*2836*	*2749*
Template per Conceptual Unit (Range)	*236 (7-666)*	*305 (6-800)*

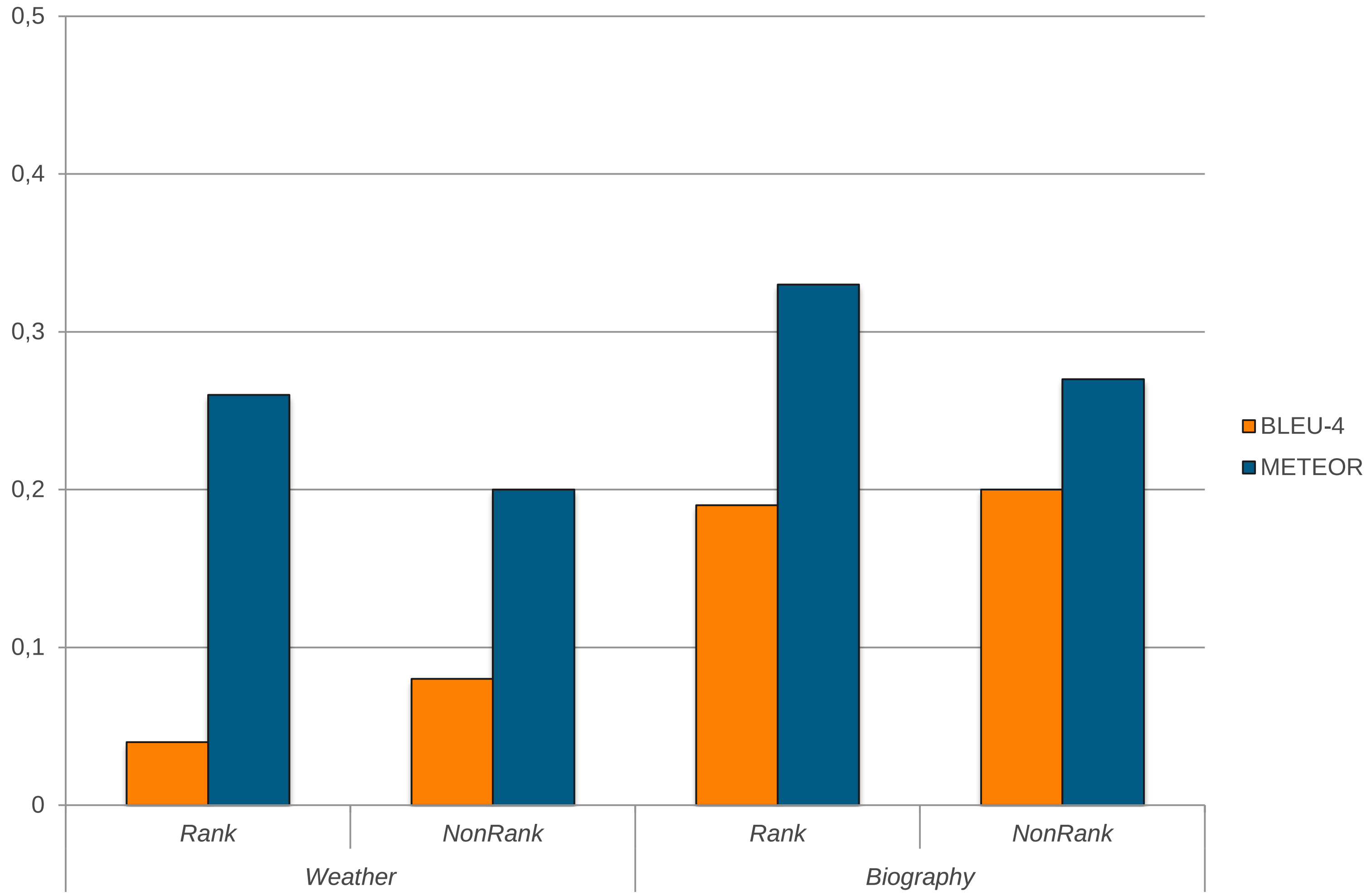
EVALUATIONS

- *Original* texts compared against *Rank* and *Non-Rank* with automatic metrics (*Biography = 350*; *Weather* = 209):
  - *BLEU-4* – 4-gram overlap
  - *METEOR* – unigram weighted f-score less penalty based on chunking dissimilarity
  - *Syntactic Variability* – percentage of unique template sequences across all documents
    - Higher value (closer to 1) indicates that documents in a collection are linguistically different
    - Lower value (closer to 0) indicates that documents in a collection are linguistically similarly

Automatic Evaluations – *Syntactic Variability*



Automatic Evaluations – BLEU-4 & METEOR



AUTOMATIC METRICS

- Variability: Rank has about the same variability as the original text.
- BLEU-4: Rank is lower than NonRank
- METEOR: Rank is higher than NonRank
- Automatic metrics BLEU-4 and METEOR are not very sensitive to Content selection and Document planning:

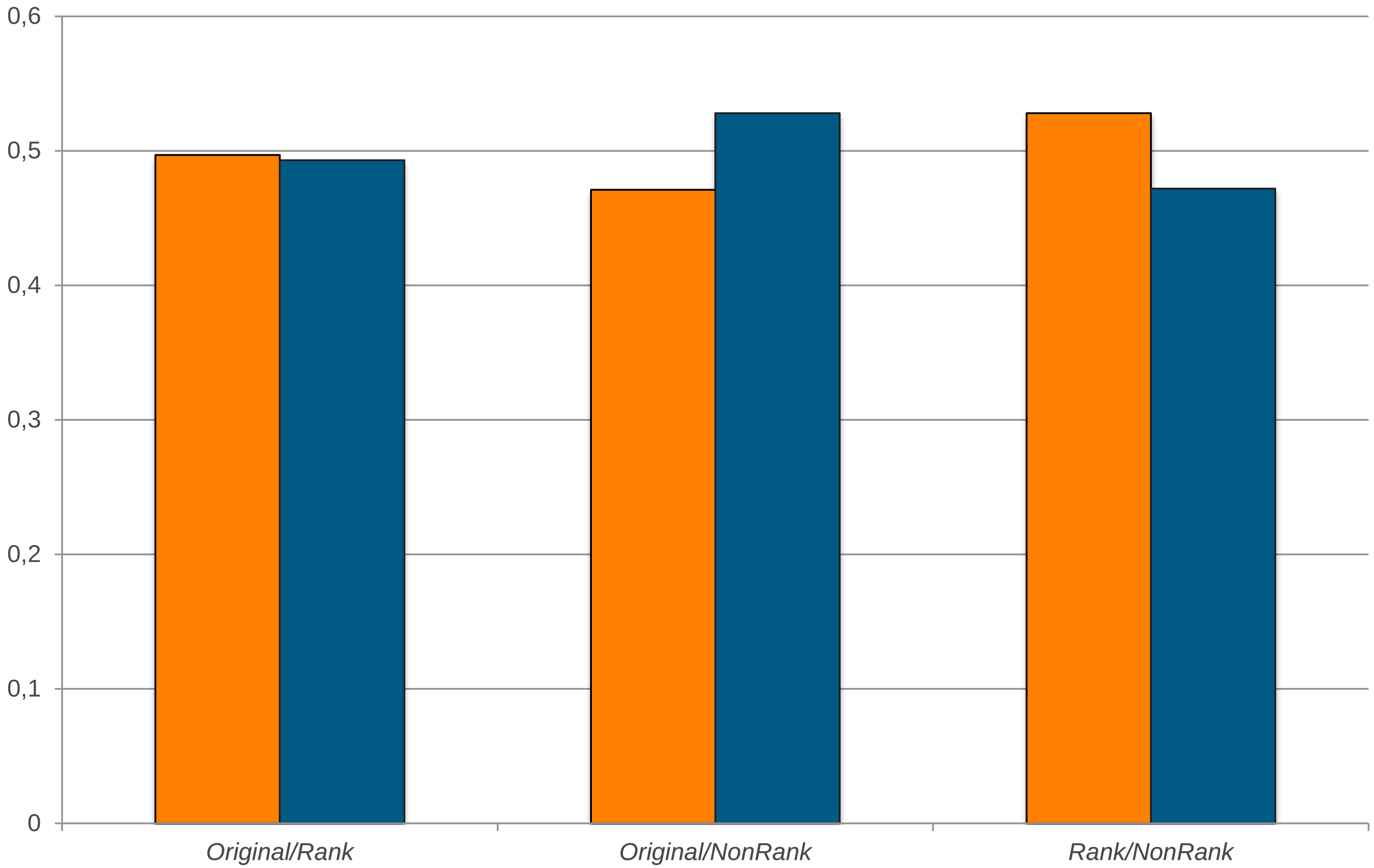
NON-EXPERT CROWDSOURCE EVALUATION

- Two tasks on Crowdflower:
  - Sentence Preference
  - Text Understandability
- Native English speakers with geographic restriction (US, UK, Australia, etc.)
- Four initial gold data responses required
  - no more than 50 responses total per person (IP address)
  - one additional gold question every four questions – had to be answered correctly to continue
- Radio buttons separated from text to avoid click bias

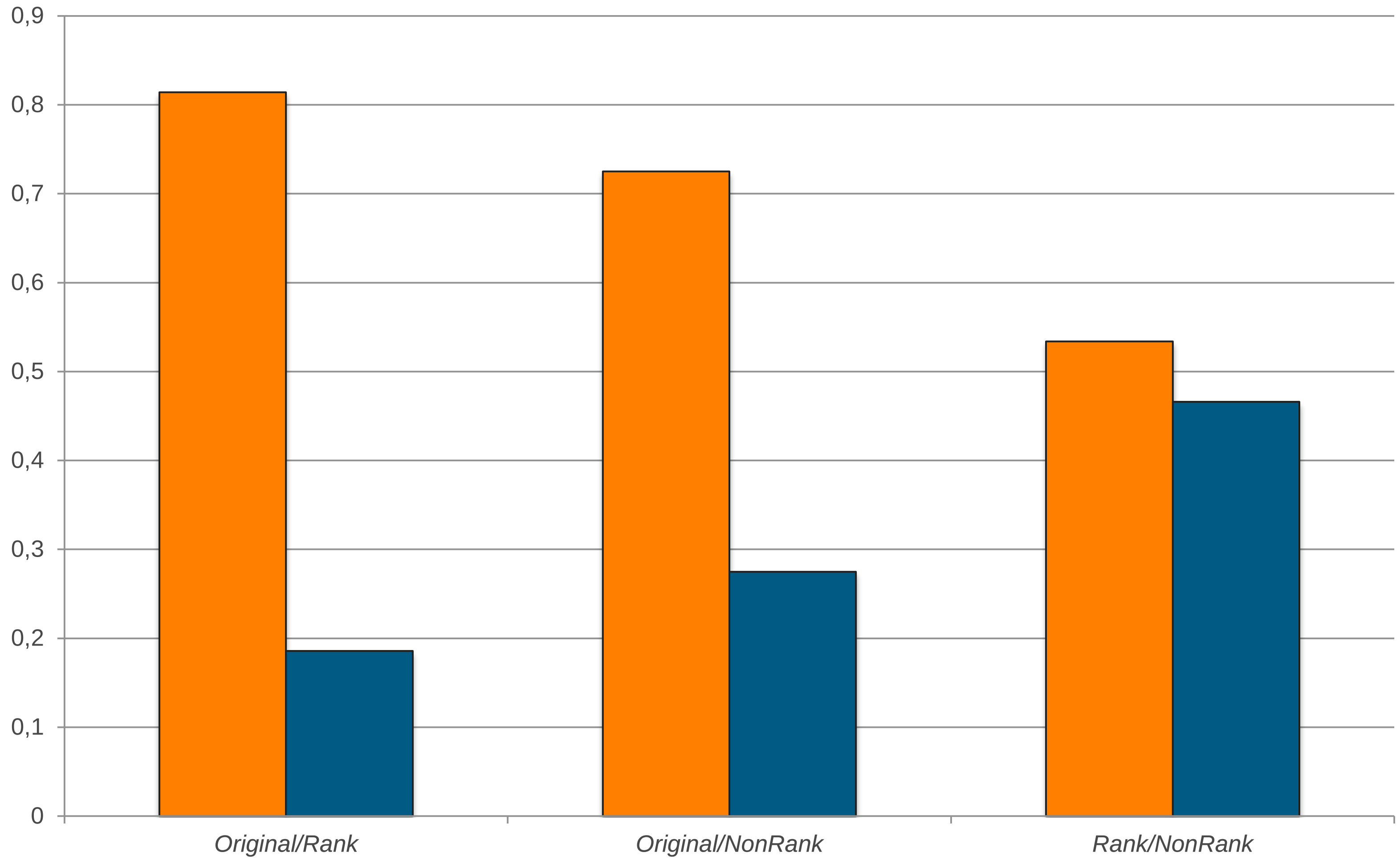
SENTENCE PREFERENCES

- You will be shown a pair of sentences expressing the same idea, for example "the cat is sitting on the mat" vs. "the cat is on the mat".
  - For each pair of sentences, indicate, as quickly as possible, which sentence you prefer.
  - Preference should be based on understandability (ease of reading, grammaticality) and informativeness (is one more informative than the other?).
- 80 sentences from *Biography,* 74 from *Weather*
- 8 judgments per sentence pair
- 3758 total judgments
- 75.87% average agreement

SENTENCE-PREFERENCE - *Weather*



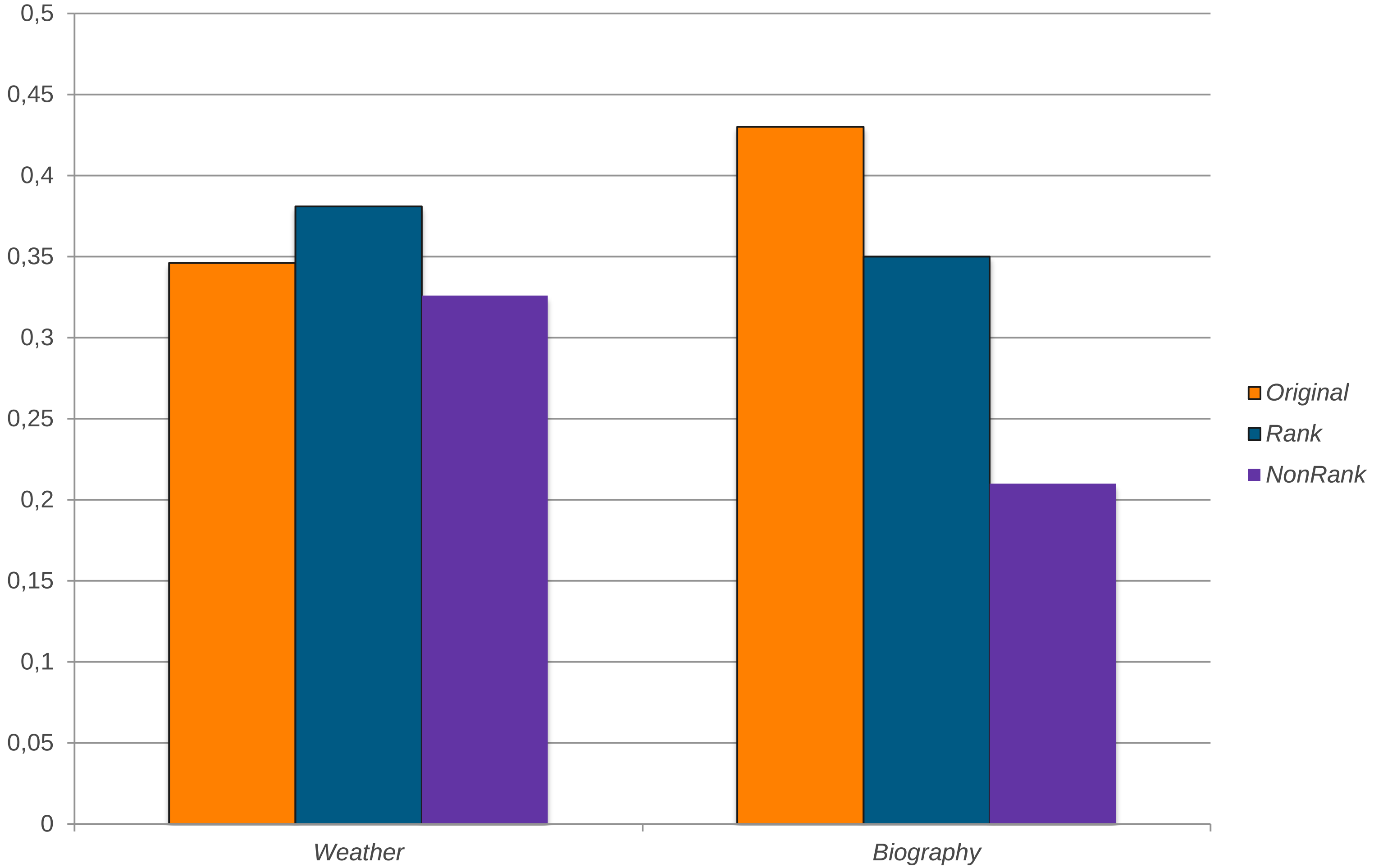
SENTENCE-PREFERENCE - *Biography*



TEXT UNDERSTANDABILITY

- Please rate the understandability of the following texts:
  - **1 = Disfluent** - Main point is not clearly understood. Severe issues with informativeness and grammar.
  - **3 = Understandable** - Main point is understood. Few issues with informativeness and grammar.
  - **5 = Fluent** - Created by a native speaker and experienced writer. Appropriately informative with no grammatical mistakes.
- 120 texts per domain (240 total)
- 8 judgments per text
- 1920 total judgments
- 69.51% average agreement

TEXT UNDERSTANDABILITY– FLUENT (“5”) RATINGS



THOMSON REUTERS

EXPERT *BIOGRAPHY* EVALUATION

- Sentence-Preference
  - 3 judgments per sentence (76.22% agreement)
  - Similar trend as the non-expert crowd
    - *Original* preferred over *Rank* and *NonRank*
  - But, *NonRank* preferred 70% to the *Rank*'s (30%)
- Text-Understandability
  - 3 judgments per document (72.95% agreement)
  - Similar trend as the non-expert crowd
    - *Original* had a higher fluency than the *Rank* and *NonRank*
  - But, *NonRank* had 10% higher "Fluent" rating (58.22%) compared to the *Rank* (47.97%)
- Why? *NonRank* generations are shorter and more concise – in keeping with editorial standards,
  - Note that training data didn't always follow this guideline.

CONCLUSIONS AND DISCUSSION

- Conclusions
  - NLG generation technique:
    - New NLG framework combining learning templates and selecting content/document structure via a Ranking SVM
    - Framework is domain adaptable
  - Evaluation:
    - New automatic evaluation metric for syntactic variability
    - Crowd-source evaluation
      - showed advantage of Ranking approach for overall fluency for biography data
      - Indicated problems with domain-specific language for the weather reports
    - Expert evaluation
      - provided feedback on preferred style for biography data
- Next Steps
  - Address data consumption and its relation to coreference generation
  - Automatic template generation

Thank You!

Questions?

frank.schilder@thomsonreuters.com

http://labs.thomsonreuters.com/about-rd-careers/