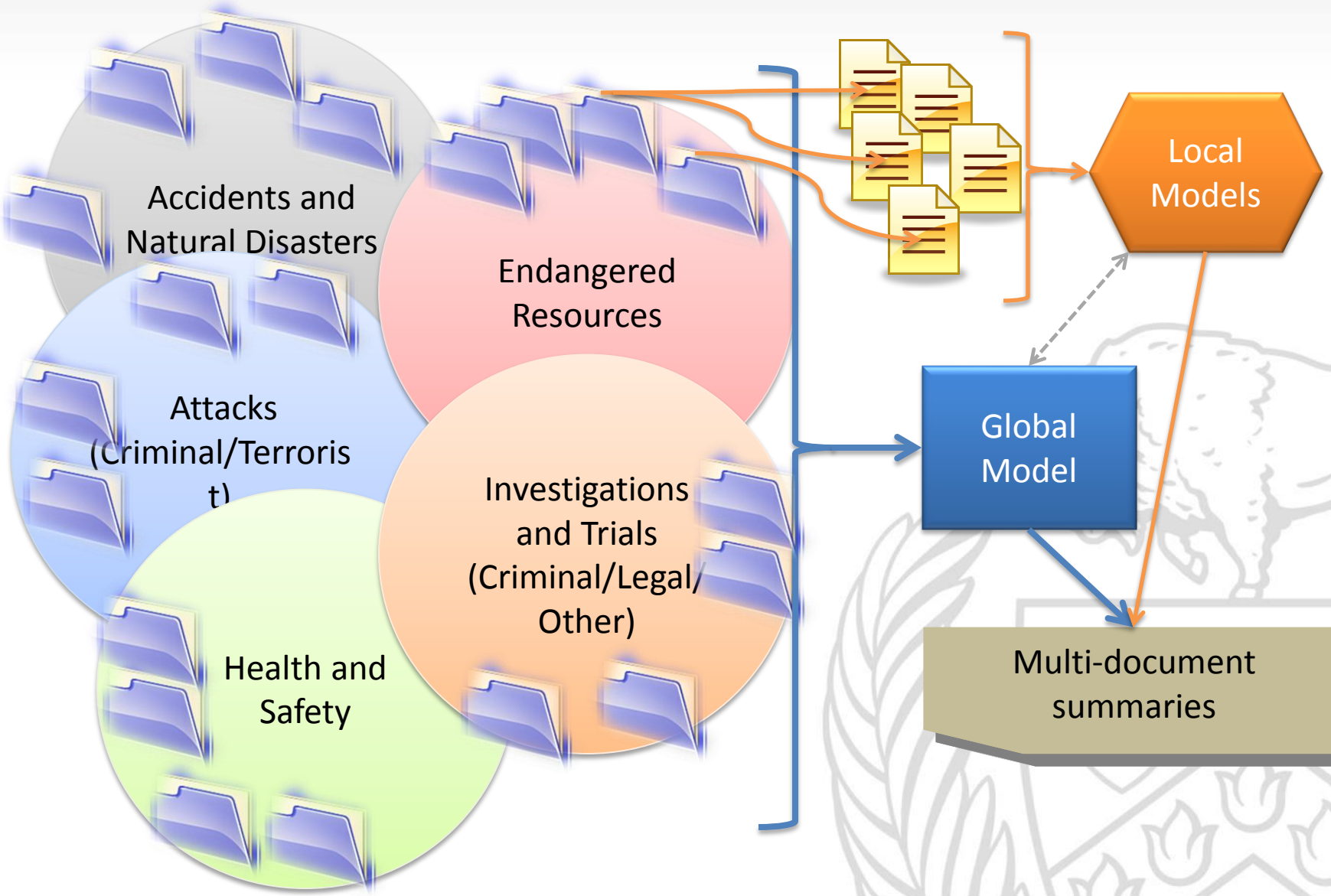
Global and Local Models for Multi-document Summarization

Pradipto Das and Rohini Srihari

SUNY Buffalo

TAC 2011, Gaithersburg, MD

Global and Local Models



An Example of a Global Model

Topics over words

Topic	Translation	Topic	Translation	Topic	Translation
सुनामी, भूकंप, चाइल, पिचिलेमू, गये, चेतावनी, खबर, शहर	Tsunami, earthquake, Chile, Pichilemu, gone, warning , news, city	विमान, एयर, फ़्रांस, जहाज़, ब्राज़ील, ए,४४७, गायब, महासागर, फ़्रांसीसी	flight, Air, France, Brazil, A, 447, disappear, ocean France	चीन, ओलंपिक, बीजिंग, गोर, समारोह, स्वर्ण, स्टेडियम, खेलों	China, Olympic, Beijing, Gore, function, stadium, games

Topics over controlled vocabulary

Topic	Translation	Topic	Translation	Topic	Translation
सुनामी, भूकंप, भूकंप:xx->xx, शहर, स्थानीय, यू०टी०सी०, मेयर, सुनामी:xx->xx	Tsunami, earthquake, earthquake:xx->xx, city, local, UTC, Mayor, Tsunami:xx->xx	ब्राज़ील, ए, गायब, खोज, उड़ान, विमान:xx->xx, महासागर, जहाज़:xx->xx, एयर:xx->xx, हवाई, क्षेत्र	Brazil, A, disappeared, search, flight, aircraft:xx->xx, ocean, ship:xx->xx, air:xx->xx, air, space	चीन,ओलंपिक चीन:xx->xx, बीजिंग, ओलंपिक:xx->xx, गोर:xx->xx, गोर, स्वर्ण, बीजिंग:xx->xx, नेशनल	China, Olympic, China:xx->xx, Olympic:xx->xx, Gore:xx->xx, Gore, gold, Beijing:xx->xx, National

Bi-Perspective Document Structure

Words in Para 1

Words in Para 2

National Institute of Standards and Technology

From Wikipedia, the free encyclopedia

"NIST" redirects here. For other uses, see NIST (disambiguation).

The **National Institute of Standards and Technology** (**NIST**), known between 1901 and 1988 as the **National Bureau of Standards** (**NBS**), is a measurement standards laboratory which is a non-regulatory agency of the United States Department of Commerce. The institute's official mission is to:[1]

> Promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

National Institute of Standards and Technology

NIST

Agency overview

Headquarters	Gaithersburg, Maryla
Annual budget	US$820 million (200
	US$662 million (est.
	US$722 million (est.

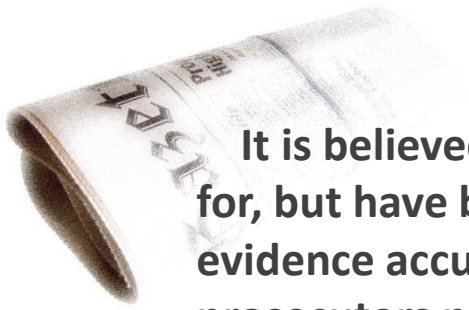
Categories: Standards organizations | National Institute of Standards and Technology | Gaithersburg, Maryland | United States Department of Commerce agencies | Government agencies established in 1901

Manually edited Wiki category tags – words that summarize/ categorize the document

Wikipedia

Understanding the Two Perspectives

- Imagine browsing over reports in a topic cluster



It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr. Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year.

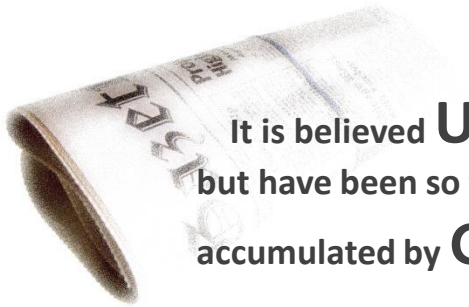
This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms. Holland.

Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.

News Article

Understanding the Two Perspectives

- What words **can we remember** after a first browse?



It is believed **US investigators** have asked for, but have been so far refused access to, evidence accumulated by **German prosecutors** probing allegations that former **GM** director, **Mr. Lopez**, stole industrial secrets from the **US** group and took them with him when he joined VW last year.

This **investigation** was launched by **US** President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of **Ms. Holland**.

Dorothea Holland, until four months ago was the only **prosecuting** lawyer on the **German** case.

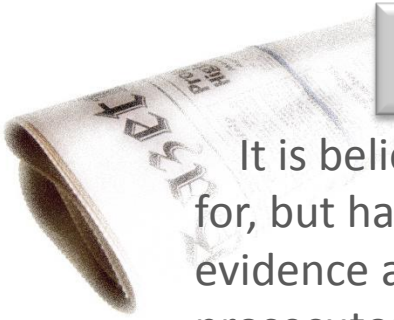
German, US, investigations, GM, Dorothea Holland, Lopez, prosecute

The "document level" perspective

News Article

Understanding the Two Perspectives

❑ What helped us generate the Document Level perspective?



The "word level" perspective

It is believed **US investigators** have asked for, but have been so far refused access to, evidence accumulated by **German** prosecutors probing allegations that former **GM** director, **Mr. Lopez**, **stole** industrial **secrets** from the **US** group and took them with him when he joined **VW** last year.

This **investigation** was launched by **US** President **Bill Clinton** and is in principle a far more simple or at least more single-minded pursuit than that of **Ms. Holland**.

Dorothea Holland, until four months ago was the only prosecuting lawyer on the **German** case.

Named Entities

- LOCATION
- MISC
- ORGANIZATION
- PERSON

Important Verbs and Dependents

- WHAT HAPPENED?

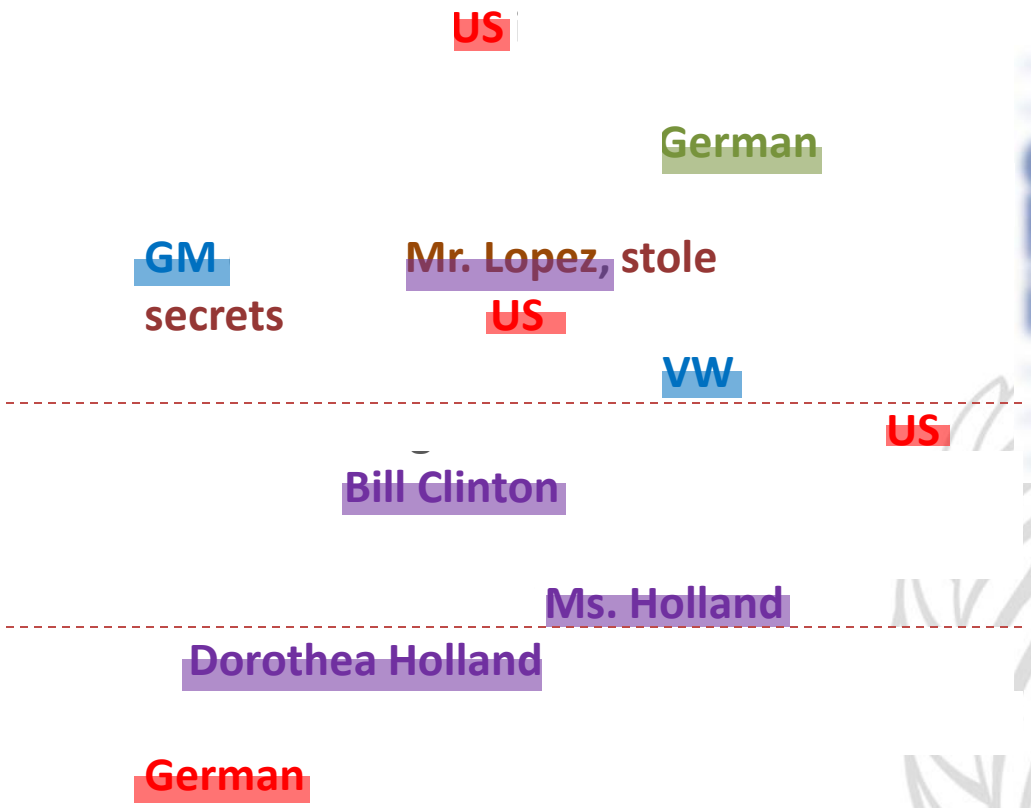
German, US, investigations, GM, Dorothea Holland, Lopez, prosecute

The "document level" perspective

News Article

What if we turn the document off?

- Summarization power of the perspectives



German, US, investigations, GM, Dorothea Holland, Lopez, prosecute

Assumptions of the Global Models

- Documents are at least tagged from two different perspectives – either implicit or explicit and **one perspective affects the other**
  - Simplest example of implicit WL tagging – binned positions indicating sections
  - Simplest example of implicit DL tagging – tag cloud

Begin (0)	It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr. Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year.
Middle (1)	This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms. Holland.
End (2)	Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.

Document Level Perspectives

• Guided Summarization Track

Centers of Attentions (with regard to grammatical or semantic roles)	Menu_foods:ne->ne, pet:nn->nn, unit:nn->nn, Henderson:ne->ne, wheat:nn->nn, food:subj->nn etc.
Top 20 $(\text{tf-idf})_{\text{docset}}$ words + Top 5 most frequent non-stopwords in the documents	Menu_Foods, pet, associate, plant, sell, source, FDA, Henderson, agency, shelf, test, unit, Canadian, dog, food etc.

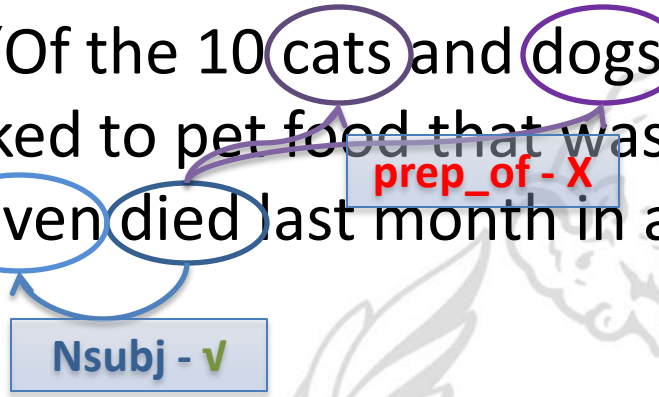
• Multilingual Track

Centers of Attentions (without regard to grammatical or semantic roles)	उत्तर(North):xx->xx, घायल(injured):xx->xx, जांच(investigation):xx->xx, लंदन(London):xx->xx, पुलिस(police):xx->xx etc.
Top 20 $(\text{tf-idf})_{\text{docset}}$ words + Top 5 most frequent non-stopwords in the documents	जांच(investigation), घरों(houses), तलाशी(search), पुलिस(police), स्टेशन(station), किंग्स(King's), क्रॉस(Cross), हमले(attack)etc.

Multilingual stopwords found by Google translate

Word Level Perspectives

- **Guided Summarization Track**
  - Named Entity classes (Person, Organization, Location, Misc, Date/time/money/number/ordinal/percent)
  - Subjective class e.g. "Of the 10 cats and dogs whose deaths have been linked to pet food that was recalled over the weekend, seven died last month in a taste test conducted by…"

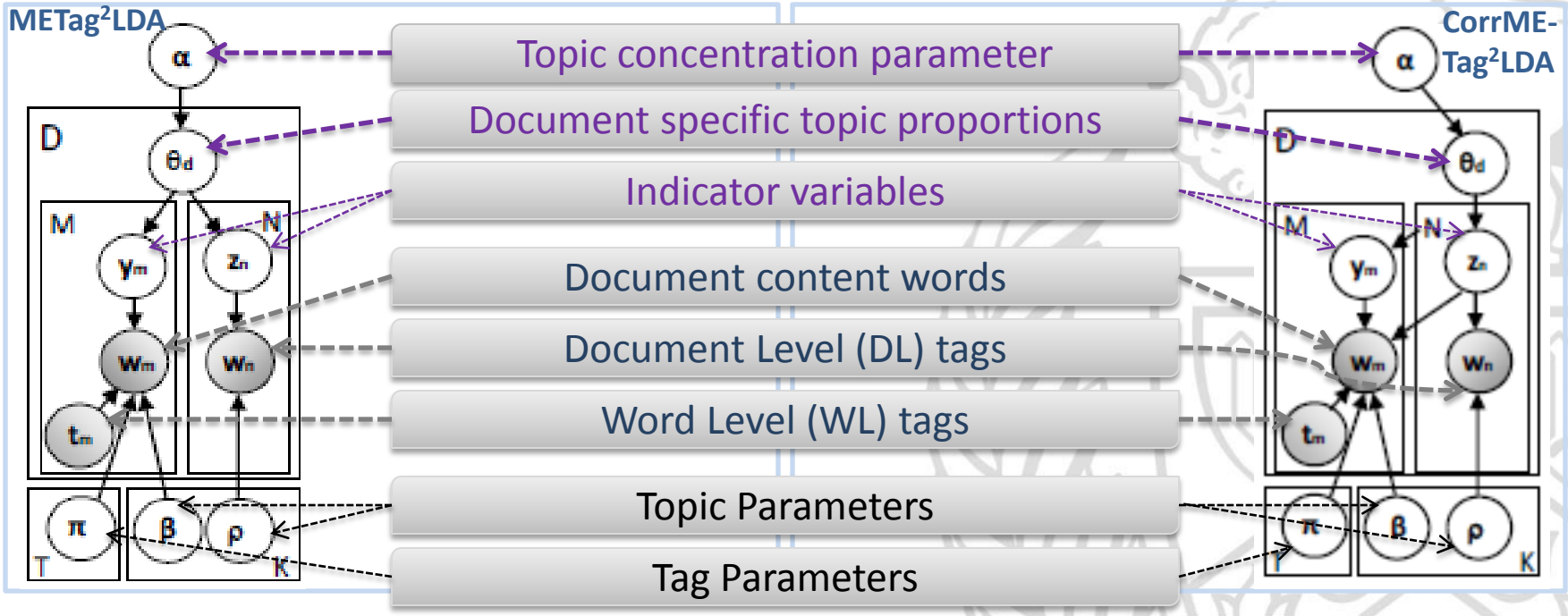


- **Multilingual Track**
  - {0, 1, 2, 3, 4}: Words annotated by positional bins – document segregated into 5 "sections"

Global(Background) Models

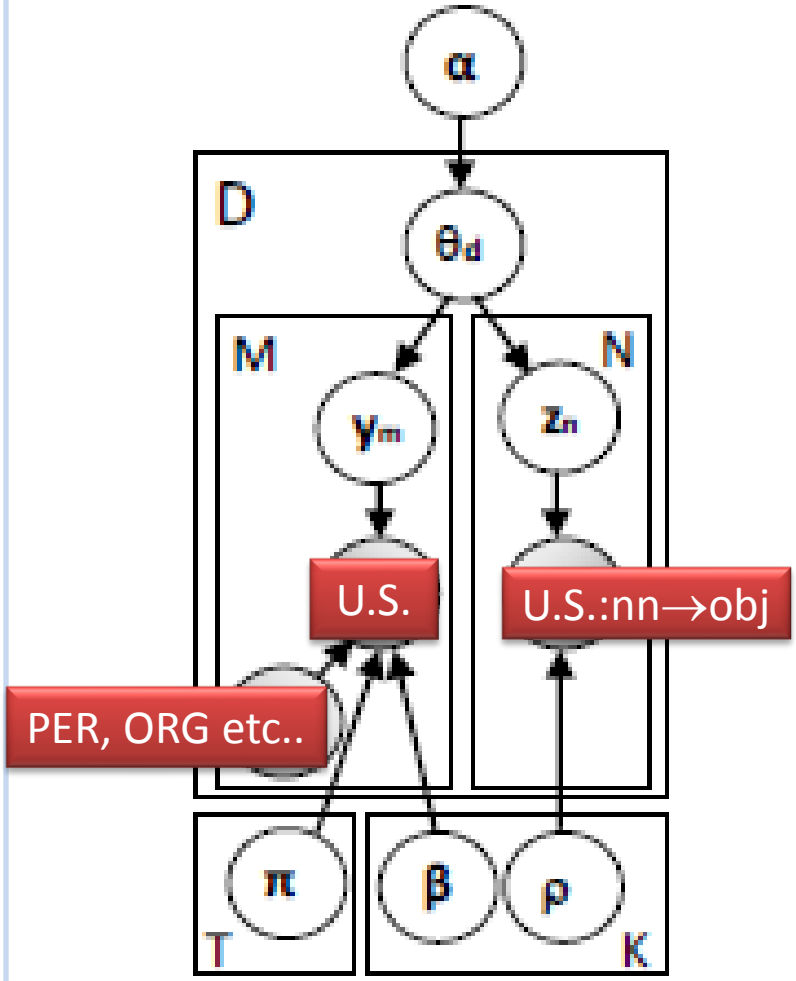
- METag^{2}LDA: **A topic** generating **all DL tags** in a document **doesn't necessarily mean** that the same topic generates all words in the document
- CorrMETag^{2}LDA: **A topic** generating ***all* DL tags** in a document **does mean** that the same topic generates all words in the document

The idea was to assign weights to words in sentences from a generative standpoint

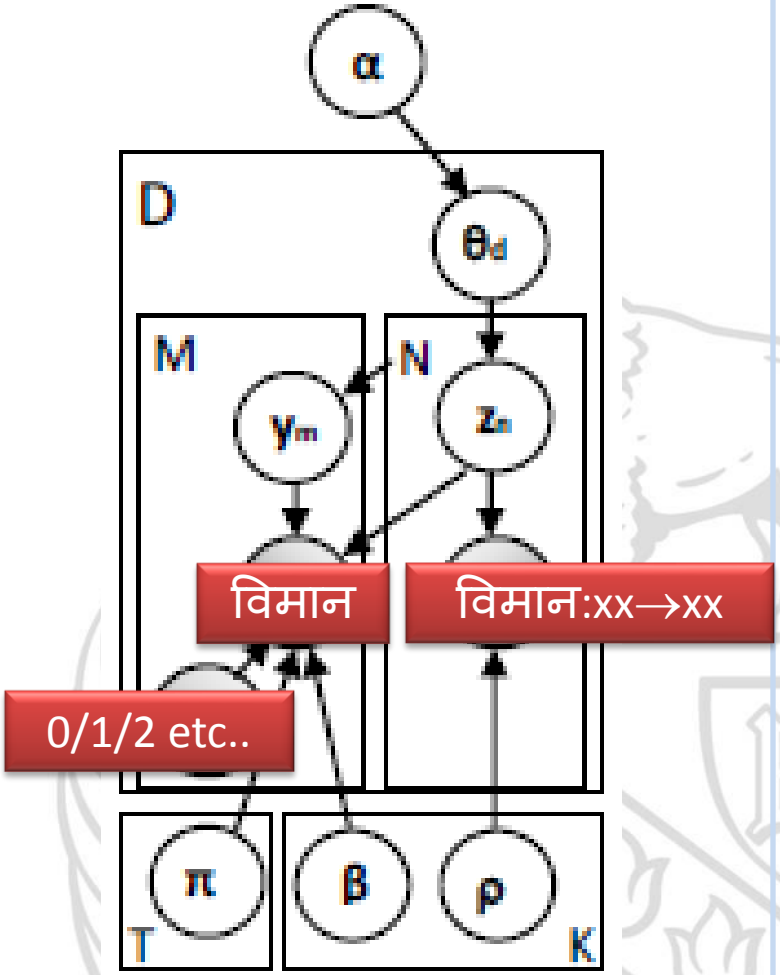


Global(Background) Models

METag²LDA



CorrMETag²LDA



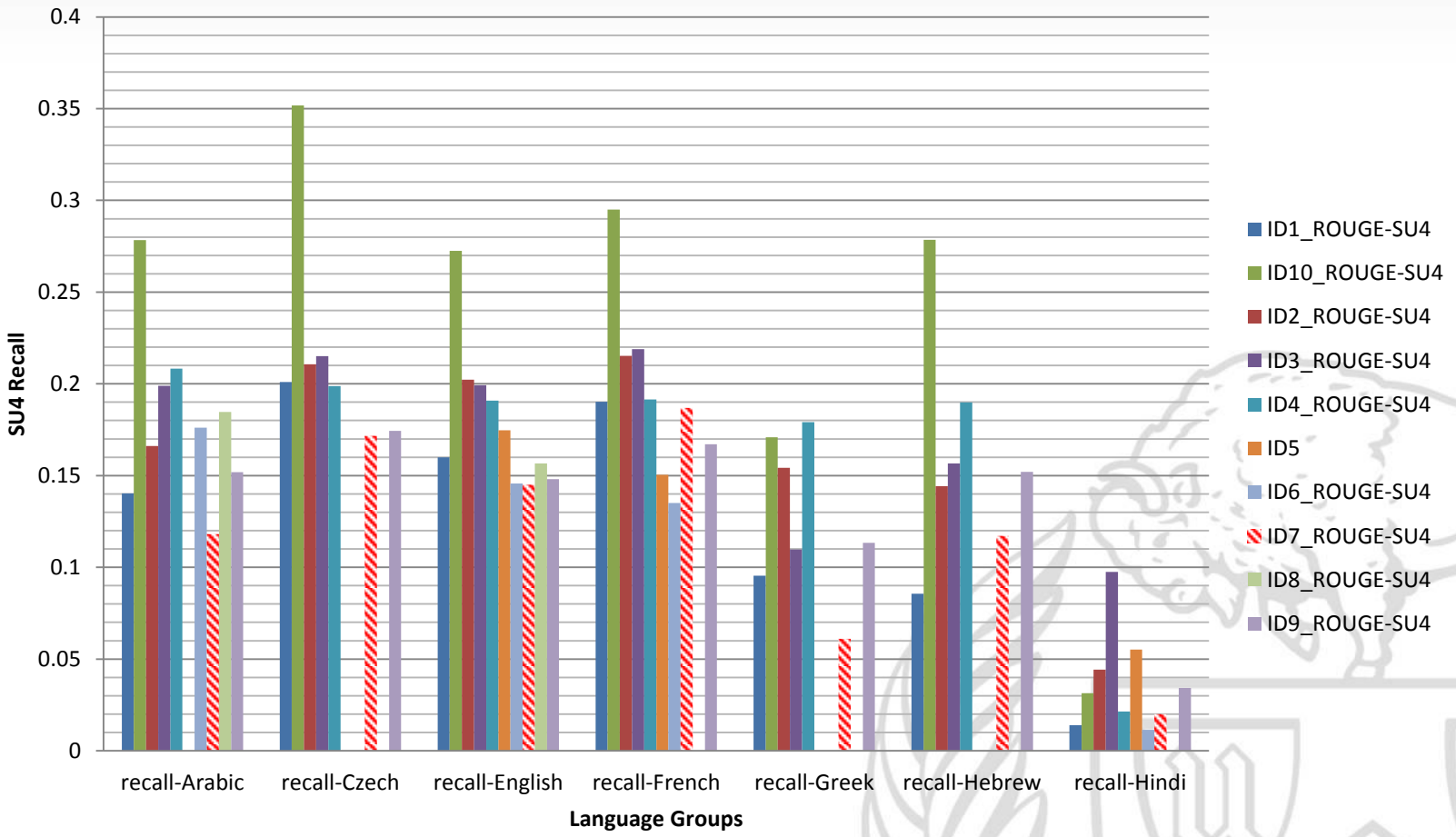
Local Models - Guided

- **Guided Summarization Track**
  - Collection of a bag of all nouns (**Bag-nn**) which are not proper nouns from the Document Level perspective
  - Collection of a bag of all verbs (**Bag-vb**) which are not stopwords from the Word Level perspective
  - Collection of the dependency parsing (using open-source Stanford CoreNLP parser) outputs for each sentence in the docset
- $score_{global}(q, sentence_i) = \sum_{j=1}^{N_{q(i)}} \sum_{k=1}^{latent_{topics}} \boldsymbol{p}(\boldsymbol{w_{q(i)}}, \boldsymbol{z_k})$ where *sentence*$_i$ is within a context that is fit to the model METag^{2}LDA
- Finally $score_{sentence_i}$ by greedily optimizing: $F = score_{global+local}(\boldsymbol{query}, sentence_i) - redundancy(sentence_i, sentence_j) - redundancy(sentence_i, prev_{summary})\delta\{1, has_{prev_{summary}}\}$
- Overlapping sentence removal and heuristic sentence pruning afterwards

Local Model - Multilingual

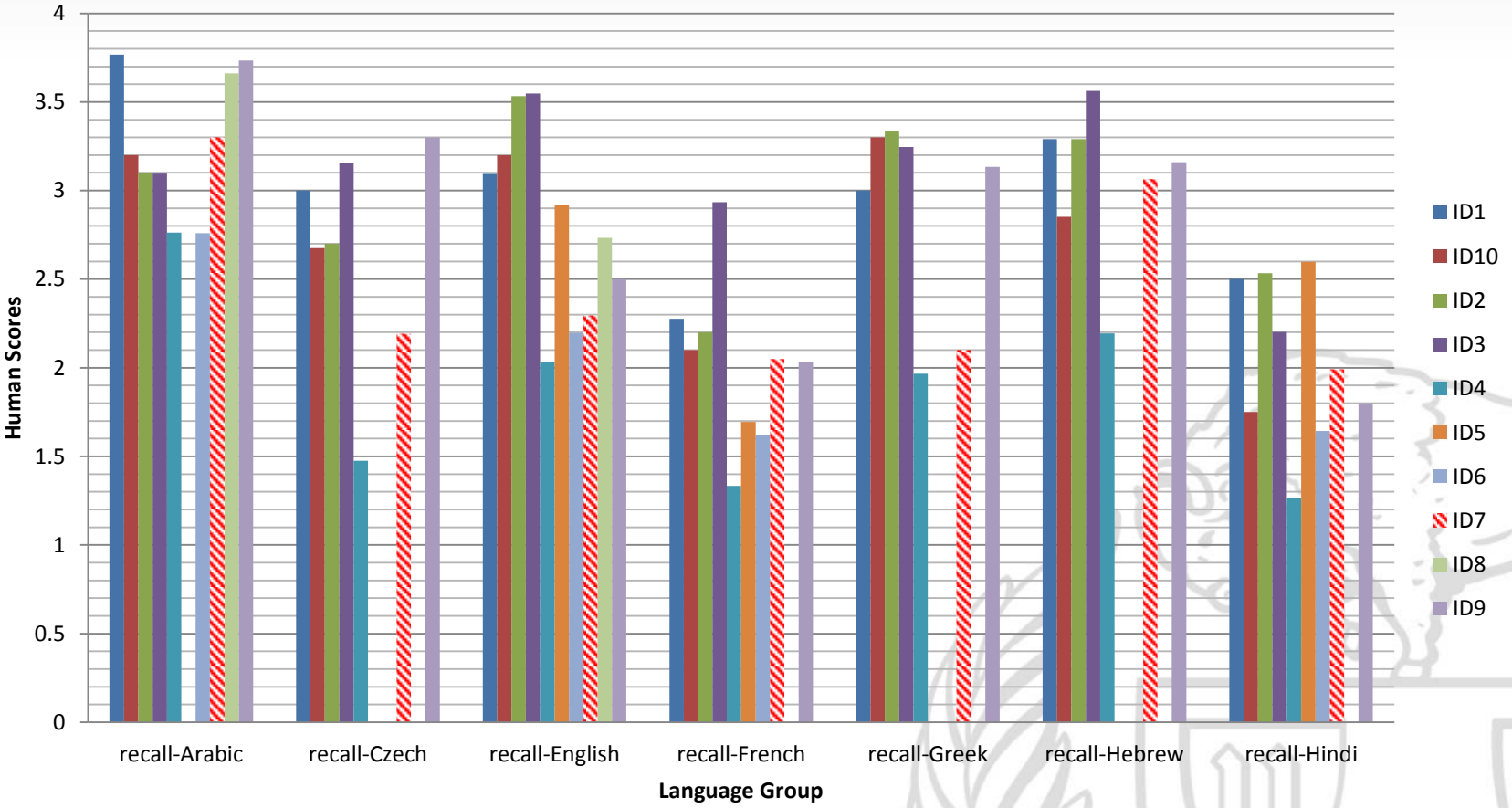
- **Multilingual Track**
  - Purely based on probabilities!
    - Take a multilingual sentence context (s±1) where the central sentence is at least 10 words long
    - Obtain the log likelihoods **of the contexts** to the trained corrMETag^{2}LDA model with 30 topics
    - **Order the sentences in descending order of likelihoods**
    - **Post-processing only involves keeping sentences within a length threshold, checking for overlaps and removing sentences beginning with a quote**

Multilingual Track: Automatic scoring



Our system ID: 7 – Not doing well by just fitting sentence contexts likelihoods

Multilingual Track: Manual Scoring



Our system ID: 7 – Doing average by just fitting sentence contexts likelihoods

But scores are stable across most languages

Conclusions

- Experiment with Sum_{CF} as in Nenkova et. al., 2006 over all "important" sentential words not just query words
- Improve and add simple but effective local models that uses closed class words like stopwords
- Verb identification in multilingual documents can help local models