



# Linking Content in Unstructured Sources

Marie-Francine Moens

Department of Computer Science - LIIR

Katholieke Universiteit Leuven, Belgium

<http://www.cs.kuleuven.be/~liir/>

WWW 2010 April 27, 2010

# Overview

---

1. Motivation and introduction to the applications
2. Introductory concepts
3. Monolingual linking of content
4. Cross-lingual linking of content
5. Cross-media linking of content
6. Conclusions and future perspectives

---

# 1. Motivation and introduction to the applications

# Problem definition

---

- **Unstructured sources**
  - Unstructured: presupposes that although the semantic information in the source is not *immediately* computationally transparent, it can nevertheless be retrieved by taking into account surface regularities
  - Sources = digital content: **natural language statements, images, video**, audio, gestures, etc. and their combinations
- **Linking** and making associations are primordial in human perception: an **intelligent machine** should be able to do so



# Story understanding

## The Three Little Pigs

Once upon a time there were three little pigs and the time came for them to leave home and seek their fortunes.

Before **they** left, their mother told them " Whatever you do , do it the best that you can because that's the way to get along in the world.



The first little pig built his house out of straw because it was the easiest thing to do.

The second little pig built his house out of sticks. This was a little bit stronger than a straw house.

The third little pig built his house out of bricks.

One night the big bad wolf, **who** dearly loved to eat fat little piggies, came along and saw the first little pig in **his** house of straw. **He** said "Let me in, Let me in, little pig and I'll blow your house in!"



# RAS

## Regulation Assistance System

Version 1.0

[main](#)  
[settings](#)

[lookup](#)  
[definitions](#)

Displaying provision: 40.cfr.279.40.a.4

### Definitions

(4) This subpart does not apply to transportation of used oil from household do-it-yourselfers to a regulated used oil generator, collection center, aggregation point, processor/refiner, or burner subject to the requirements of this part. Except as otherwise provided in this section, the term used oil means any oil that is collected or used at any of the following facilities where household do-it-yourselfer used oil is collected: 40.cfr.279.40.a.1 40.cfr.279.40.a.2 40.cfr.279.40.a.3

Used oil generator means any person, by site, whose act or process produces used oil or whose act first causes used oil to become subject to regulation.

### Links to References

#### Suggested search terms:

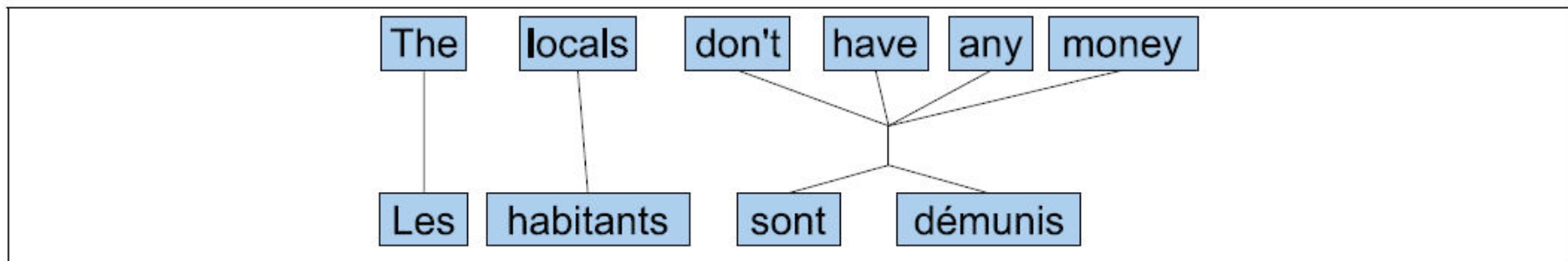
[aggregation points](#) [collection center](#) [household do-it-yourselfer](#) [oil generator](#)

### Search Terms/ Concepts

Internet

## Alignment in statistical machine translation

---



An alignment between an English and a French sentence, in which there is a many-to-many alignment between English and French words:  
Needs phrase alignment.

[Jurafsky & Martin Chapter 25 2006] [TermWise & WebInsight projects]

After the latest Fed rate cut, stocks rose across the board.  
 Winners strongly outpaced losers after Greenspan cut interest rates again.

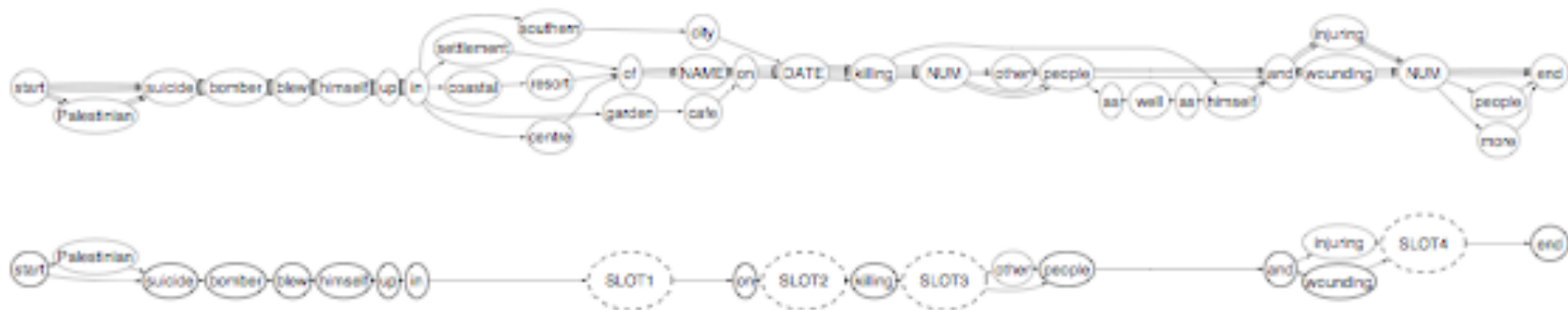


Figure 3: Lattice and slotted lattice for the five sentences from Figure 2. Punctuation and articles removed for clarity.

[Barzilay & Lee HLT-NAACL 2008] [DAISY Stevin project]



# Linking names and faces

[Labeled faces in the wild dataset]



U.S. President **George W. Bush** (2nd R) speaks to the press following a meeting with the Interagency Team on Iraq at Camp David in Maryland, June 12, 2006. Pictured with **Bush** are (L-R) Vice President **Dick Cheney**, Defense Secretary **Donald Rumsfeld** and Secretary of State **Condoleezza Rice**.

[Pham, Moens & Tuytelaars IEEE T Multimedia 2010] [IWT-SBO AMASS++project]

# Ambiguous names

## Tom Mitchell

Fredkin Professor of AI and Machine Learning  
Chair, [Machine Learning Department](#)  
[School of Computer Science](#)  
Carnegie Mellon University

412-268-2611, [Tom.Mitchell@cmu.edu](mailto:Tom.Mitchell@cmu.edu), [Resume](#), [A personal interview](#)

Assistant: [Sharon Cavlovich](#), 412 268-5196

What is Machine Learning, and where is it headed?



Tom Mitchell

### Tom Mitchell

**Research Fellow**  
Vulnerability and Poverty Reduction Team  
Climate Change and Disasters Group  
CV (Word)

**Tel:** 44 (0)1273 915757

**E-mail:** [t.mitchell@ids.ac.uk](mailto:t.mitchell@ids.ac.uk)

**Administrative contact:** Hannah Bywaters  
([h.bywaters@ids.ac.uk](mailto:h.bywaters@ids.ac.uk))



### Biography

Dr. Tom Mitchell is a Research Fellow at IDS, having previously been a member of the Benfield UCL Hazard Research Centre. He specialises in climate change adaptation and disaster risk reduction. His interests include pro-poor climate and disaster governance and he co-ordinates the children in a changing climate programme.

[Angheluta & Moens ECIR 2007]

WWW 2010

10

# Information mashup

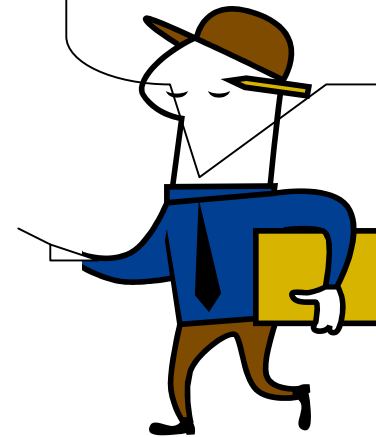


[Gyllstrom & Moens SIGIR 2010] [EU FP7 PuppyIR project]

## Multimodal linking



Go to that white car !



[http://upload.wikimedia.org/wikipedia/commons/b/b5/Toy\\_robot.JPG](http://upload.wikimedia.org/wikipedia/commons/b/b5/Toy_robot.JPG)



- 
- The focus in this tutorial = **linking equivalent content** useful to:
    - Search, reasoning with information, summarization, building of cross-lingual and cross-media dictionaries, ...
    - Because equivalent content comes in many forms:
      - Can be seen as a **translation problem**

- 
- Besides equivalence relation: other interesting “discourse” relations that often signal complementary information [Allan PhD thesis 1994] ...:
    - Revision Cf. Automatic hypertext generation
    - Summary and expansion
    - Comparison - contrast
    - Tangent - aggregate, ....
  - Or “event” relations: e.g., who, did what to whom where when ...  
Cf. Relation extraction and event template filling

# Problem definition

---

- How can we **automatically realize this linking**?
- Are there **generic** algorithms?
- Can we **reduce human supervision**?
- When using an interlingua, can we realize the linking jointly **without** a separate translation of source and target into an **interlingua**?

Our methods our data-driven

**Australian  
Open won by  
Maria  
Sharapova**

**Victory!**

**Location:  
Melbourne  
Park**

**Sports**

**Tennis 2008**

Sharapova beats Ivanovic to win Australian Open

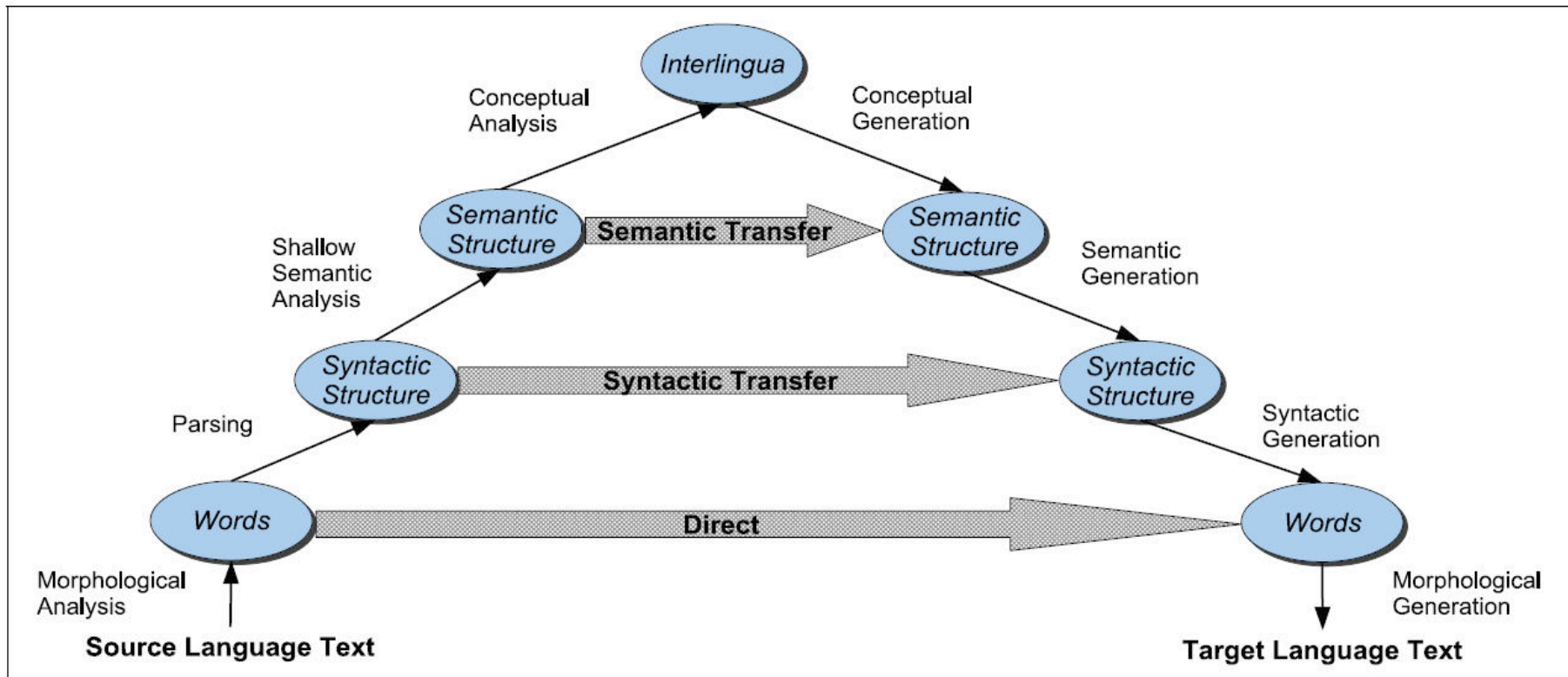
A year after being on the wrong end of one of Russian didn't drop a set in seven matches at Melbourne Park, including wins the most-lopsided losses in a Grand Slam final, Sharapova wrapped up her third major title with a 7-5, 6-3 victory over fourth-seeded Ana Ivanovic on Saturday. The 20-year-old over three of the top four ranked players, erasing 12 months worth of painful memories in the wake of her 6-1, 6-2 loss to Serena Williams last year. After Ivanovic sprayed a forehand wide on match point, Sharapova dropped to her knees and appeared to be fighting back tears as she waved and blew kisses to the crowd. Then she dropped her racket in her chair before heading to shake hands and exchange high-fives with her father and supporters.

**Content labels are valuable, for instance, for linking  
information**

**But, an almost infinitive number of ways in which information  
can be linked**

- **Problem on the data side =**
  - Huge amount of patterns that signal equivalent content
  - Ambiguous patterns
- **Problem on the usage side =**
  - If the linking is performed solely based on identified interlingua concepts:
    - Almost infinite set of concepts: the meta-language risks to become as complex as natural language
    - Interlingua concepts do not contain anymore their low level features and contexts

# Vauquois Triangle



[Jurafsky & Martin Chapter 25 2006]

# Maybe complementary ...

---

- Linking of **lower level features**: more flexible (cf. success of classical search engines)
- **Interlingua** can be useful as **latent class** when linking data

# Extra difficulties

---

- Seldom parallel information: **null links**
- Cardinality ratio:
  - **1:1, 1:N, N:1 or N:M**
- Problem of segmentation
- Information can be hierarchically organized e.g., parts that make up a whole
- Often asymmetry, though results can be symmetrized later
- Typical for unstructured sources:
  - **Similar form, different meaning (polysemy)**
  - **Different form, same meaning (synonymy)**
- **Watch for computational complexity !**



- 
- There exist already a lot of research on content linking in different disciplines:
    - Natural language processing
    - Computer vision
    - Data mining
    - Cross-media mining
    - ...
  - Surprisingly: **many similarities between the algorithms** (although often independently developed)

---

## **2. Introductory concepts**

# Some definitions

---

- **Cross-modal**: coming from multiple information sources, which consist of multiple types of content, i.e. multimedia content
- **Parallel corpus**: corpus with parallel information: might be texts, where one text is exact translation in another language of the other
- **Comparable corpus**: corpus with similar information, but each source might also contain different information: e.g., text with images that illustrate part of the info in the text

- 
- It is all about **finding similarities** !
  - But, we deal seldom with exact matches (other media, other languages, even other language patterns in monolingual context, ...)
  - And, we deal with heterogeneous feature patterns: different media, languages

=>

- Finding associations, correlations and possibly clustering of information
- Sometimes we need to detect auxiliary latent classes in the data

- 
- Let us start with an example from alignment or linking of multilingual content to formalize the problem:
    - best studied in the literature

# Alignment - linking models

Ma soeur aime ses chiens bruns.

And my sister loves her brown dogs.

**Alignment  $a_i$**  is defined as a subset of the Cartesian product of word positions:  $a_i \subseteq \{(y, x) : y = 1, \dots, Y; x = 1, \dots, X\}$

Associations  $y \rightarrow x = a_{ix}$

If  $a_{ix} = 0$ , alignment with the “empty” or NULL object

A parallel/comparable corpus consisting of  $S$  object pairs:

$\{(s_s, t_s) = 1, \dots, S\}$  with corresponding alignments  $\mathbf{a}$

# Alignment - linking models

---

## Three main steps:

- Find an appropriate model  $M$  for the linking of source and target – **Modeling**
- Estimate parameters of the model  $M$ , e.g. from empirical data – **Parameter estimation**
- Find the optimal linking according to the model  $M$  and its parameters – **Linking/Alignment recovery**

# Parameter estimation

---

## Two main approaches:

- **Association approach** – Alignment is based on similarity and association measures
- **Latent class approach** - Parameters are modeled as hidden parameters in a statistical translation model



# Association approach

- Utilize a function of similarity between the candidate pairs
- A huge variety of scoring functions:
  - Cosinus, Dice coefficient, pointwise mutual information statistic, chi-square, t-score, log-likelihood measure, kernel functions, tree kernels, string kernels, ...
  - Similarity function might be learned from training data

## Rescuers comb debris for survivors after Mississippi tornado

Rescue crews in Mississippi continued to search for survivors Sunday from a powerful tornado that ripped through the state a day earlier, killing 10 people, injuring dozens of others and leveling scores of homes. f



**Yazoo City, Mississippi (CNN)** -- Massive cleanup efforts got under way Monday after several tornadoes ripped through the South, killing at least 12 people -- 10 in Mississippi -- and leaving a swath of devastation in the region, from Louisiana to Alabama.

[Manning & Schütze 1999]  
[Moschitti ECML 2006]  
[Bhattacharya & Getoor TKDD 2007]

...

# Association approach

---

- In one way or another the functions deal with incomplete matches and additional constraints can be modeled (e.g., forbidden links)
- Best suited for: 1:1 correspondences
- The results can be clustered: yields group based associations

# But ...

---

- We often deal with **events with uncertain outcomes** (foreign word has many candidate translations, a face can be linked to many candidate names, ...)
- **Probability distribution** = function that maps possible outcomes to values between 0 and 1:
  - We might model an event with a standard distribution: e.g., uniform, binomial or normal
  - We might collect statistics about the event and estimate the distribution by maximum likelihood estimation
  - We can also model more complex distributions such as joint or conditional distributions for related events

- 
- We often learn from **incomplete data**:
    - We do not know what the alignments or links are in the data, or have only a few “cognates” (i.e., links that we are sure and have to learn the other ones)
    - But in large data set we assume that links are redundant
    - One way to address this problem: **Expectation Maximization (EM) algorithm**:
      - Iteratively:
        - Computes the probability of possible links
        - Collects counts
        - Builds an improved model based on these counts

# Latent class approach

---

- Generative models: often treat **alignment as a hidden process**
- The unknown parameters  $\theta$  are determined by maximizing the likelihood of the alignments on the training corpus (e.g., by using an EM algorithm):

$$\hat{\theta} = \arg \max_{\theta} \prod_{s=1}^S \sum_{\mathbf{a}} p_{\theta}(t_s, \mathbf{a} | s_s)$$

[e.g., Och & Ney CL 2003]

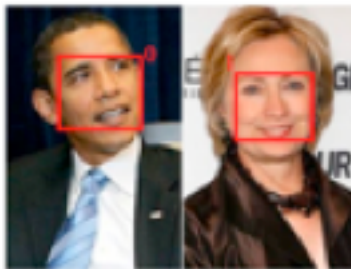
- 
- Although for a given object with candidate pairs there are a large number of alignments  $a_j$ , we can always find the alignment with highest probability:

$$\hat{a}_i = \arg \max_{a_i} p_{\theta}(t_i, a_i | s_i)$$

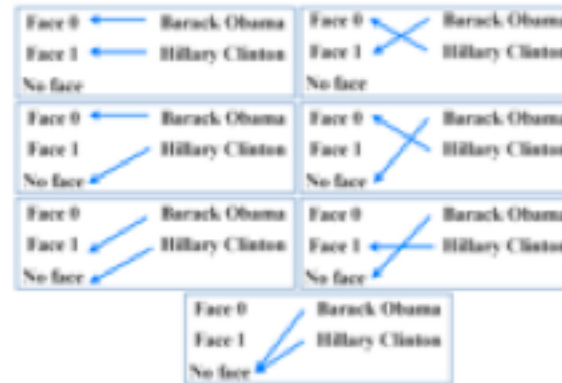
- When many possibilities to select from: Viterbi like decoding

# Hidden variables

- Alignment parameter: strength of the alignment
- “Interlingua” concept
- ...



President-elect Barack Obama is inching closer to naming former rival Sen. Hillary Clinton as his secretary of state, ABC News has learned. (Getty Images)



e.g., the weight of a possible link scheme is modeled as a hidden variable

- In classification/recognition: given inputs  $\mathbf{x}$  and their labels  $y$ :

---

  - **Generative model:** attempts to model underlying probability distributions that generate the data and learns a model of the joint probability  $p(\mathbf{x}, y)$  and then selects the most likely label: e.g.,
    - e.g., Bayesian networks, Naive Bayes, ...
    - supervised, but usually completely unsupervised
  - **Discriminative model:** is trained to model the conditional probability  $p(y|\mathbf{x})$  directly and selects the most likely label  $y$ , or learns a direct map from inputs  $\mathbf{x}$  to the labels: e.g.,
    - maximum entropy model, support vector machine
    - often trained in a supervised way



# Often graph problems

---

- Graph cuts for clustering
- Random walks in graphs
- Inference in Bayesian networks and undirected networks
- ...

# Evaluation

---

- Comparison with ground truth data, manually built
- Individual recognitions of links:

$$recall = \frac{|crl|}{|cl|} \qquad precision = \frac{|crl|}{|rl|}$$

where  $crl$  represents the set of correctly recognized links,  $cl$  the set of correct links and  $rl$  the set of recognized links

- Grouped recognitions: B-cubed precision, B-Cubed recall computed for each mention  $m_i$  given ground truth cluster  $M_{m_i}$  and machine-generated cluster  $C_{m_i}$  to which  $m_i$  belongs:

# Evaluation

---

$$precision_{mi} = \frac{|C_{mi} \cap M_{mi}|}{|C_{mi}|}$$

$$recall_{mi} = \frac{|C_{mi} \cap M_{mi}|}{|M_{mi}|}$$

$$precision = \frac{1}{|\mathbf{m}|} \sum_{mi \in \mathbf{m}} precision_{mi}$$

$$recall = \frac{1}{|\mathbf{m}|} \sum_{mi \in \mathbf{m}} recall_{mi}$$

where  $\mathbf{m}$  = set of mentions to be grouped

- F-measure: combines recall and precision

$$F = \frac{(\beta^2 + 1) \text{ precision} \times \text{ recall}}{\beta^2 \text{ precision} + \text{ recall}}$$

where  $\beta$  = a factor (=1, harmonic mean) that indicates the relative importance of recall and precision

# Evaluation

---

- Alignment-Link error rate:

$$AER(Su, Po; A) = \frac{|A \cap Su| + |A \cap Po|}{|A| + |Su|}$$

- Gold standard with sure (*Su*) alignment point and possible (*Po*) alignment points

[See also Fraser & Marcu Comp. Ling 2007]

We cite some results to illustrate the capabilities of the techniques, we do not describe the experimental setup, but refer to the cited papers for additional results and their details.

---

# **3. Monolingual linking of content**

# Linking of content in text sources

---

- Most studied: linking of entity mentions:
  - **In one document**: anaphora resolution and noun phrase coreference resolution includes mention clustering and disambiguation
  - **Across documents**: noun phrase coreference resolution includes mention clustering and disambiguation

# Arizona governor signs immigration bill

By the CNN Wire Staff

April 23, 2010 10:16 p.m. EDT



## STORY HIGHLIGHTS

- Executive order requires training on implementing law without racial profiling
- Measure "threatens to undermine basic notions of

Phoenix, Arizona (CNN) -- Arizona Gov. Jan Brewer signed a bill Friday that requires police in her state to determine whether a person is in the United States legally, which critics say will foster racial profiling but supporters say will crack down on illegal immigration.

[\[www.cnn.com\]](http://www.cnn.com)

---

## ■ Coreference resolution:

- Task of grouping all mentions  $m_i$  of entities in a document (news story, related Webpages) into equivalent classes so that all the mentions in a given class refer to the same discourse entity (for simplicity we refer to the mentions by their syntactic head)
- Number of equivalence classes is not specified in advance, but bounded by the number of mentions



Typical features in a single-document noun phrase coreference resolution task of the syntactic heads,  $m_i$  and  $m_j$ , of two candidate coreferent noun phrases in text  $T$  where  $m_i < m_j$  in terms of word position in  $T$ .

FEATURE	VALUE TYPE	VALUE
Number agreement	Boolean	True if $m_i$ and $m_j$ agree in number; False otherwise.
Gender agreement	Boolean	True if $m_i$ and $m_j$ agree in gender; False otherwise.
Alias	Boolean	True if $m_i$ is an alias of $m_j$ or vice versa; False otherwise.
Weak alias	Boolean	True if $m_i$ is a substring of $m_j$ or vice versa; False otherwise.
POS match	Boolean	True if the POS tag of $m_i$ and $m_j$ match; False otherwise.
Pronoun $m_i$	Boolean	True if $m_i$ is a pronoun; False otherwise.
Personal pronoun $m_j$	Boolean	True if $m_j$ is a personal pronoun; False otherwise.
Relative pronoun $m_i$	Boolean	True if $m_i$ is a relative pronoun; False otherwise.
<u>Anaphoricity</u> $m_j$	Boolean	True if $m_j$ is an anaphor; False otherwise.
Appositive	Boolean	True if $m_j$ is the appositive of $m_i$ ; False otherwise.



Definiteness	Boolean	True if $m_i$ is preceded by the article “the” or a demonstrative pronoun; False otherwise.
Grammatical role	Boolean	True if the grammatical role of $m_i$ and $m_j$ match; False otherwise.
Proper names	Boolean	True if $m_i$ and $m_j$ are both proper names; False otherwise.
Named entity class	Boolean	True if $m_i$ and $m_j$ have the same semantic class (e.g., person, company, location); False otherwise.
WordNet feature	Boolean	True if sense of $m_i$ and is synonym, antonym or hypernym of any sense of $m_j$ ; False otherwise.
Modifier match	Boolean	True if $m_i$ and $m_j$ share the same modifier; False otherwise.
Discourse distance	Integer $\geq 0$	Number of sentences or words that $m_i$ and $m_j$ are apart.



[Moens IRS 2006] [Bengtson & Roth EMNLP 2009]

# Pairwise coreference classifier

---

- $\mathbf{m}$  is a set of mentions ( $m_i$ ) (e.g., noun phrases) in the document
- $\mathbf{x}$  is the set of pairs of noun phrases:  $x_{ij} = \{m_i, m_j\}$
- $\mathbf{y}$  is the set of variables representing each pairwise coreference decision  $y_{ij}$  involving mentions  $m_i$  and  $m_j$
- Binary random variable  $y_{ij} = 1$  if  $m_i$  and  $m_j$  are coreferent
- Let  $F = \{x_{ij}, y_{ij}\}$  be a set of feature functions over  $x_{ij}$  (e.g., Boolean)
- $p(y_{ij} | x_{ij})$ : computed with a classification model

# Pairwise coreference classifier

---

- Maximum entropy model (i.e., (multinomial) logistic regression):

$$p(y_{ij}|x_{ij}) = \frac{1}{Z} \exp\left(\sum_k w_l f_l(x_{ij}, y_{ij})\right), \quad 0 < w_l < \infty$$

$f_l(x_{ij}, y_{ij})$  = one of the  $k$  binary-valued feature functions

$w_l$  = real-valued weight parameter estimated from the training data

$Z$  = normalizing constant

# Constructing the clusters

---

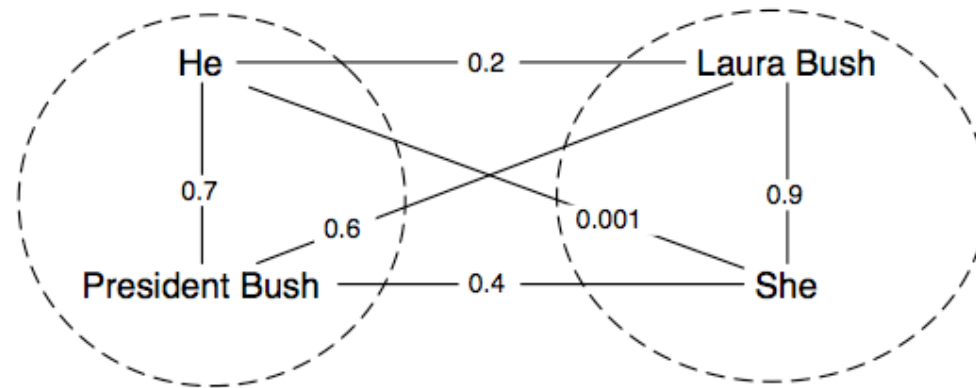


Figure 1: An example noun coreference graph in which vertices are noun phrases and edge weights are proportional to the probability that the two nouns are coreferent. Partitioning such a graph into disjoint clusters corresponds to performing coreference resolution on the noun phrases.

# Constructing the clusters

---

- Goal: partition graph into clusters with high intra-cluster edge weights and low inter-cluster edge weights:
  - Based on strength (or probability) of detected relations
  - Possibly augmented with heuristic constraints of forbidden merge of mentions
- Often greedy clustering: each noun phrase  $m_j$  is assigned to the same cluster as the closest preceding noun phrase  $m_i$  for which  $p(y_{ij}|x_{ij}) > \delta$  (threshold) (e.g.,  $\delta = 0.5$ )

MUC Precision	MUC Recall	MUC F
82.7	69.9	75.8

Table 5: Evaluation of our system on unseen Test Data using MUC score.

	Precision	Recall	$B^3$ F
Culotta et al.	86.7	73.2	79.3
<b>Current Work</b>	88.3	74.5	<b>80.8</b>

Table 4: Evaluation on unseen Test Data using  $B^3$  score. Shows that our system outperforms the advanced system of Culotta et al. The improvement is statistically significant at the  $p = 0.05$  level according to a non-parametric bootstrapping percentile test.

[Bengtson & Roth EMNLP 2008]

# Enforcing transitivity with ILP

- After computing the pairwise classification decisions:
  - Use **integer linear programming** to enforce transitivity constraints:

$$\max \sum_{x_{ij}} \log p(y_{ij}|x_{ij}) \cdot y_{ij} - \log(1 - p(y_{ij}|x_{ij})) \cdot (1 - y_{ij})$$

- $p(y_{ij}|x_{ij})$ : computed with a classification model (see above)

[Finkel & Manning ACL-HLT 2008]



# Enforcing transitivity with ILP

---

- Add binary constraints on each of the variables:  $y_{ij} \in \{0, 1\}$
- Add constraints over each triplet of mentions to enforce transitivity:  $(1 - y_{ij}) + (1 - y_{jk}) \geq (1 - y_{ik})$ 
  - ensures that whenever  $y_{ij} = y_{jk} = 1$  also  $y_{ik} = 1$
- Use ILP tool to solve the ILP optimization problem
- Solution for short text because of computational complexity of ILP

- 
- Important cues are [Haghighi & Klein EMNLP 2009]:
    - Syntactic structures that signal preferred references
    - Semantic matches or constraints
  - Although not studied: language models might help in giving evidence of whether one word might be replaced by another word in the considered context [Deschacht & Moens EMNLP 2009]
  - The problem is also studied with hidden coreference variables given the observed mentions [Haghighi & Klein ACL 2007] [Wick & McCallum Tech. Rep. 2009]

# Linking of entities across documents

---

- 2 problems:
  - **Homonymy** = names have the same writing, but refer to different entities:
    - E.g., persons disambiguation on the Web, cf. Web People Search Task (WePs)
  - **Synonymy** = names are written differently, but refer to the same entity (cf. within document noun phrase coreference resolution, but “one sense per discourse” heuristic not applicable):
    - E.g., people hide their identity in different names or names might have different writing forms

# Homonymy

## Tom Mitchell

Fredkin Professor of AI and Machine Learning  
Chair, [Machine Learning Department](#)  
[School of Computer Science](#)  
Carnegie Mellon University

412-268-2611, [Tom.Mitchell@cmu.edu](mailto:Tom.Mitchell@cmu.edu), [Resume](#), [A personal interview](#)

Assistant: [Sharon Cavlovich](#), 412 268-5196

### What is Machine Learning, and where is it headed?



### Tom Mitchell

#### Tom Mitchell

**Research Fellow**  
Vulnerability and Poverty Reduction Team  
Climate Change and Disasters Group  
CV (Word)

**Tel:** 44 (0)1273 915757

**E-mail:** [t.mitchell@ids.ac.uk](mailto:t.mitchell@ids.ac.uk)

**Administrative contact:** Hannah Bywaters  
([h.bywaters@ids.ac.uk](mailto:h.bywaters@ids.ac.uk))



#### Biography

Dr. Tom Mitchell is a Research Fellow at IDS, having previously been a member of the Benfield UCL Hazard Research Centre. He specialises in climate change adaptation and disaster risk reduction. His interests include pro-poor climate and disaster governance and he co-ordinates the children in a changing climate programme.

[Angheluta & Moens ECIR 2007]

WWW 2010

56

## Synonymy

...Woodward's source in the  
Plame scandal

... senior administration official ...

... Richard Armitage ...

# Linking of entities across documents

---

- In both cases the context is important:
  - Context determines whether two mentions refer to the same entity or to different ones
  - Context:
    - Surrounding words
    - Other entities mentioned in close vicinity
    - Other linked information

# Homonymy or ambiguous names

---

- Most simple problem: given a name only disambiguation
- Approaches:
  - Feature vector represents potential coreference relationship
  - Usually supervised:
    - Training of classifier
    - Clustering of the candidate coreference relationships possibly via graph partitioning

Typical features in a cross-document noun phrase coreference resolution task of the syntactic heads,  $m_i$  and  $m_j$  of two candidate coreferent noun phrases where  $m_i$  and  $m_j$  occur in different documents.

FEATURE	TYPE	VALUE
Context word	Boolean or real value between 0 and 1	True if the context word $k$ occurs in the context of $m_i$ and $m_j$ ; False otherwise; If a real value is used, it indicates the weight of the context word; Proper names, time and location expressions in the context might receive a high weight.
Named entity class	Boolean	True if $m_i$ and $m_j$ have the same semantic class (e.g., person, company, location); False otherwise.
Semantic role	Boolean	True if the semantic role of $m_i$ matches the semantic role of $m_j$ ; False otherwise.

[Moens IRS 2006]



# Synonymy

- Given a name, find synonyms AND also disambiguation
- String edit distances might not be sufficient
- Graph based approaches: link based similarity measures between nodes (exploiting similarity of neighbors), e.g.,
  - Co-citation
  - SimRank, Connected-Triple, PageSim
  - Variety of random-walk methods

[Getoor & Diehl ACM SIGKDD Explorations 2005]  
[Liben-Nowell & Kleinberg JASIST 2007]

# Linking of entities across documents

---

- Generative, nonparametric Bayesian model of mentions in a document corpus, captures both within- and cross-document coreferences [[Haghighi & Klein ACL 2007](#)]

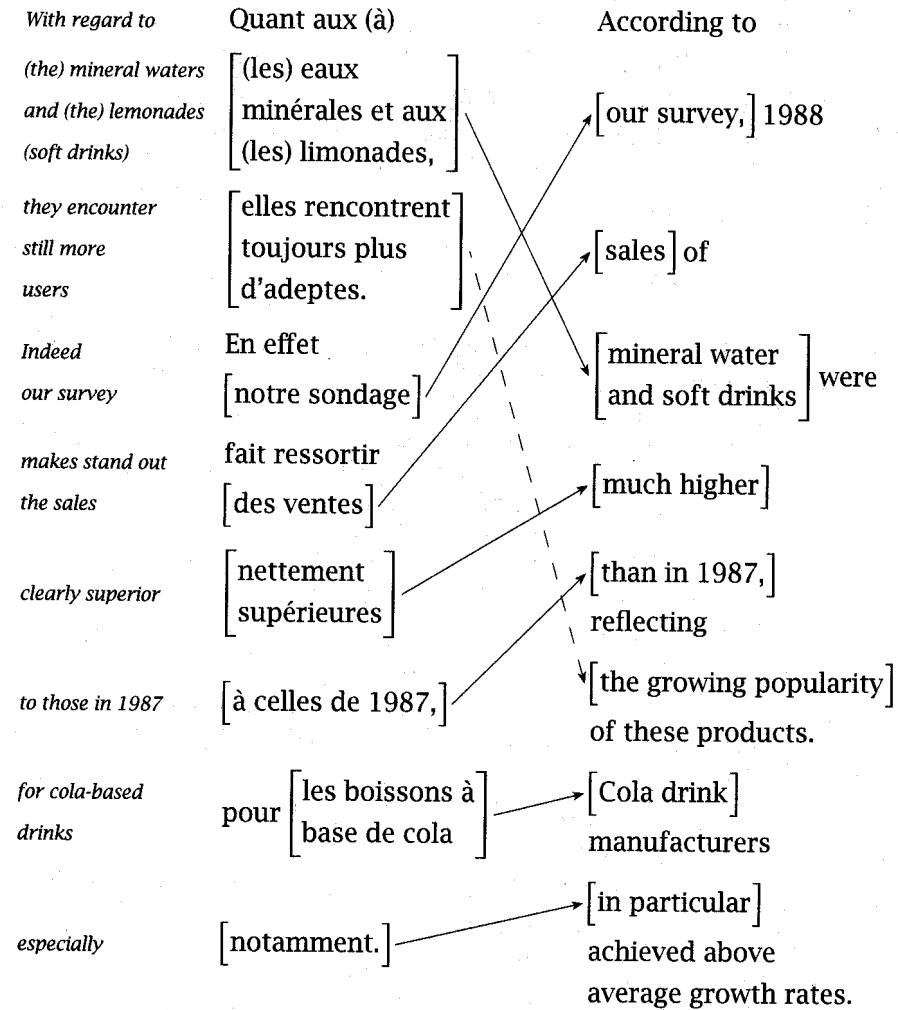
---

## **4. Cross-lingual linking of content**

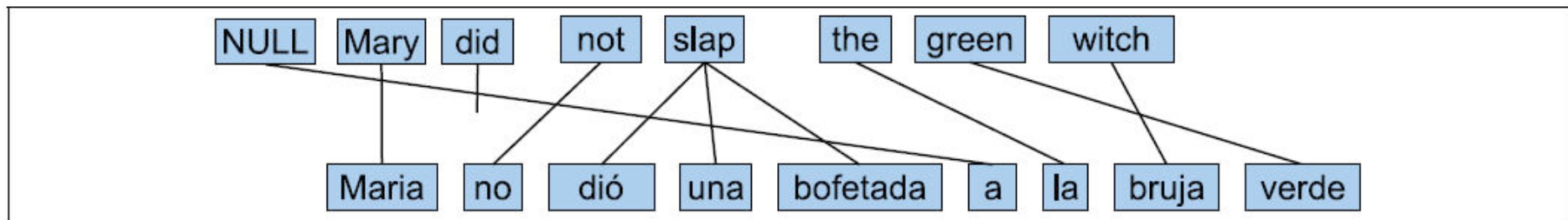
# Text alignment

---

- = identifying which text strings in one language correspond to which text strings in parallel text of other language by being the translation of each other
- Alignment of:
  - sentences and paragraphs
  - words and phrases: more difficult
- Use of statistical techniques (here illustrated with word and phrase alignment)

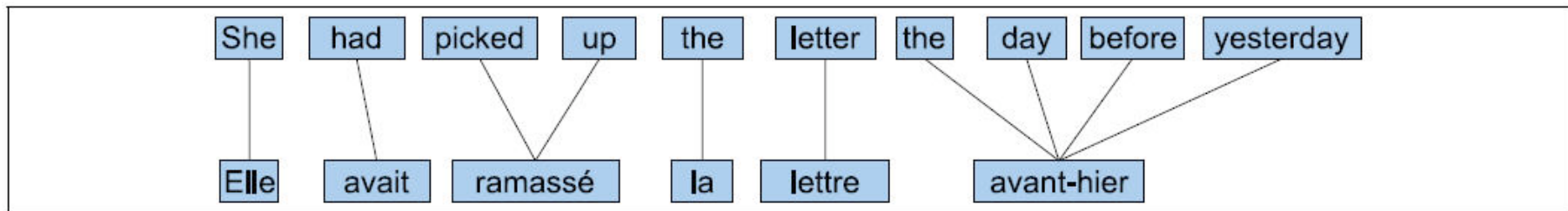


**Figure 13.2** Alignment and correspondence. The middle and right columns show the French and English versions with arrows connecting parts that can be viewed as translations of each other. The italicized text in the left column is a fairly literal translation of the French text.



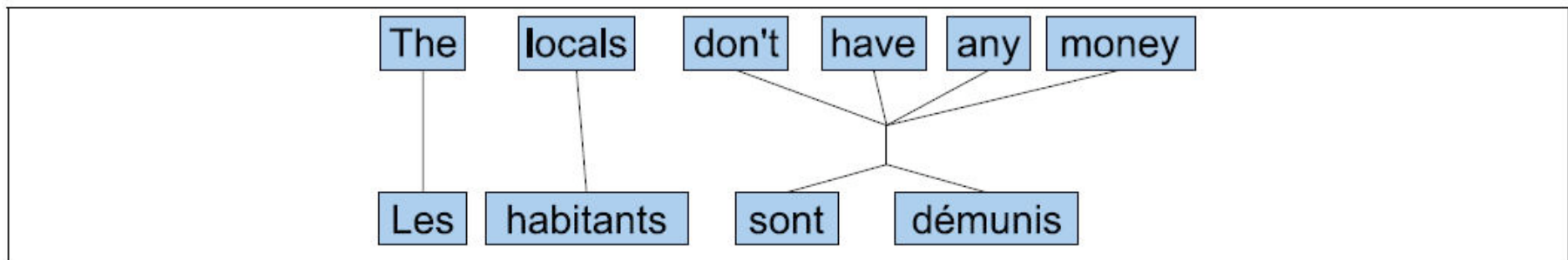
The alignment of the spurious Spanish word “a” to the English null word  $e_0$ .

[Jurafsky & Martin Chapter 25 2006]



Alignment between an English and a French sentence, in which each French word does not align to a single English word, but each English word aligns to one French word.

[Jurafsky & Martin Chapter 25 2006]



An alignment between an English and a French sentence, in which there is a many-to-many alignment between English and French words:  
Needs phrase alignment.

[Jurafsky & Martin Chapter 25 2006]



# Alignment

---

- Goals:
  - To find best alignment of sentence pair
  - To find best alignment of phrase pair
  - To find best alignment of word pair
  - => helps statistical machine translation
  - => (probabilistic) alignments can be used to build a (probabilistic) translation dictionary
- Many models:
  - Association models
  - Latent structure models: generative models IBM models 1-5, HMM model

# Association models

---

- Simpler heuristic models:
  - Word-correlation values are obtained from parallel sentences: word similarity: Dice, pointwise mutual information statistic, ...
  - Heuristics are applied to find a word alignment, often starting from highest correlating score (in 1:1 alignment)

# IBM model 1

---

- How to generate a French sentence  $\mathbf{f} = (f_1, f_2, \dots, f_J)$  from an English sentence  $\mathbf{e} = (e_1, e_2, \dots, e_I)$ ?
- The algorithm has as main steps:
  - Set the length of the French sentence
  - Choose the most probable alignment
  - Recover the French sentence from the chosen alignment

# IBM model 1

---

- Given an English sentence:  $\mathbf{e} = (e_1, \dots, e_I)$  of length  $I$  the translation probability of a French sentence  $\mathbf{f} = (f_1, \dots, f_J)$  with length  $J$  through a particular alignment  $a_j$ :

$$P(\mathbf{f}, a_i | \mathbf{e}) = \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^J t(f_j | e_{a_{ij}})$$

where  $t$  = translation probability of an English word into a French word and  $\varepsilon$  is a normalization constant

# IBM model 1

---

- Training: EM algorithm:
  - Sentence segmentation and alignment
  - Segmentation in words and training of a word aligner
  - Hidden alignment variable
  - E-step: expected counts for the  $t$  parameter
  - M-step: maximum likelihood estimate of the  $t$  probability for these counts
  - Below simple example that ignores the NULL alignment
- Decoding: finding the best alignment: Viterbi

## Two toy aligned sentences:

---

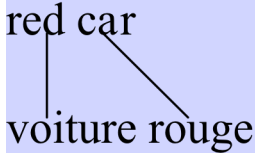
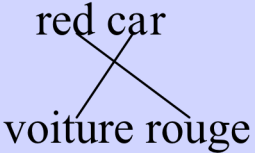
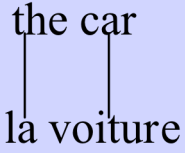
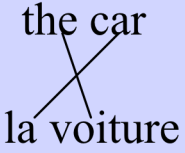
red car            the car  
voiture rouge    la voiture

Initialization: uniform probabilities:

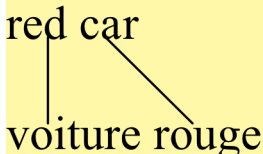
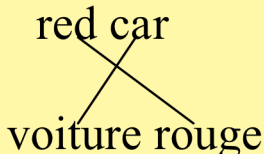
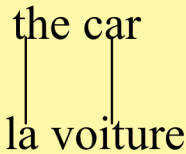
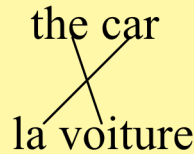
$t(\text{voiture} \text{red}) = \frac{1}{3}$	$t(\text{rouge} \text{red}) = \frac{1}{3}$	$t(\text{la} \text{red}) = \frac{1}{3}$
$t(\text{voiture} \text{car}) = \frac{1}{3}$	$t(\text{rouge} \text{car}) = \frac{1}{3}$	$t(\text{la} \text{car}) = \frac{1}{3}$
$t(\text{voiture} \text{the}) = \frac{1}{3}$	$t(\text{rouge} \text{the}) = \frac{1}{3}$	$t(\text{la} \text{the}) = \frac{1}{3}$

## E step 1:

We compute  $P(\mathbf{f}, a | \mathbf{e})$  by multiplying the  $t$  probabilities.

			
$P(\mathbf{f}, a   \mathbf{e}) = t(\text{voiture}, \text{red})$ $\times t(\text{rouge}, \text{car})$ $= \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$	$P(\mathbf{f}, a   \mathbf{e}) = t(\text{rouge}, \text{red})$ $\times t(\text{voiture}, \text{car})$ $= \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$	$P(\mathbf{f}, a   \mathbf{e}) = t(\text{la}, \text{the})$ $\times t(\text{voiture}, \text{car})$ $= \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$	$P(\mathbf{f}, a   \mathbf{e}) = t(\text{voiture}, \text{the})$ $\times t(\text{la}, \text{car})$ $= \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

We normalize  $P(\mathbf{f}, a | \mathbf{e})$  to get  $P(a | \mathbf{f}, \mathbf{e})$ .

			
$P(a   \mathbf{f}, \mathbf{e}) = \frac{1/9}{2/9} = \frac{1}{2}$	$P(a   \mathbf{f}, \mathbf{e}) = \frac{1/9}{2/9} = \frac{1}{2}$	$P(a   \mathbf{f}, \mathbf{e}) = \frac{1/9}{2/9} = \frac{1}{2}$	$P(a   \mathbf{f}, \mathbf{e}) = \frac{1/9}{2/9} = \frac{1}{2}$

---

We compute the expected fractional counts, by weighting each count by  $P(a|\mathbf{e}, \mathbf{f})$

$t(\text{voiture} \text{red}) = \frac{1}{2}$	$t(\text{rouge} \text{red}) = \frac{1}{2}$	$t(\text{la} \text{red}) = 0$
$t(\text{voiture} \text{car}) = \frac{1}{2} + \frac{1}{2}$	$t(\text{rouge} \text{car}) = \frac{1}{2}$	$t(\text{la} \text{car}) = \frac{1}{2}$
$t(\text{voiture} \text{the}) = \frac{1}{2}$	$t(\text{rouge} \text{the}) = 0$	$t(\text{la} \text{the}) = \frac{1}{2}$



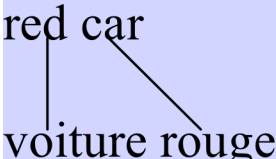
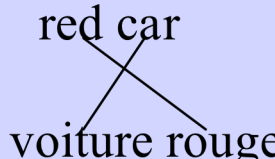
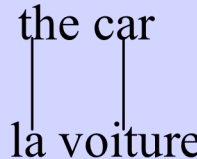

## M step 1:

We compute the MLE probability parameters by normalizing the tcounts to sum to 1.

$t(\text{voiture} \text{red}) = \frac{1/2}{1} = \frac{1}{2}$	$t(\text{rouge} \text{red}) = \frac{1/2}{1} = \frac{1}{2}$	$t(\text{la} \text{red}) = \frac{0}{1} = 0$
$t(\text{voiture} \text{car}) = \frac{1}{2}$	$t(\text{rouge} \text{car}) = \frac{1/2}{2} = \frac{1}{4}$	$t(\text{la} \text{car}) = \frac{1/2}{2} = \frac{1}{4}$
$t(\text{voiture} \text{the}) = \frac{1/2}{1} = \frac{1}{2}$	$t(\text{rouge} \text{the}) = \frac{0}{1} = 0$	$t(\text{la} \text{the}) = \frac{1/2}{1} = \frac{1}{2}$

## E step 2:

We recompute  $P(\mathbf{f}, \mathbf{a} | \mathbf{e})$  by multiplying the  $t$  probabilities. Note that the two correct alignments are now higher in probability than the two incorrect alignments.

			
$P(\mathbf{f}, \mathbf{a}   \mathbf{e}) = t(\text{voiture}, \text{red})$ $\times t(\text{rouge}, \text{car})$ $= \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	$P(\mathbf{f}, \mathbf{a}   \mathbf{e}) = t(\text{rouge}, \text{red})$ $\times t(\text{voiture}, \text{car})$ $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$P(\mathbf{f}, \mathbf{a}   \mathbf{e}) = t(\text{la}, \text{the})$ $\times t(\text{voiture}, \text{car})$ $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$P(\mathbf{f}, \mathbf{a}   \mathbf{e}) = t(\text{voiture}, \text{the})$ $\times t(\text{la}, \text{car})$ $= \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$

M step 2: ... until the values of the alignment parameters do not change anymore substantially.

# Symmetrizing alignments for identifying phrases

---

- We train English-to-French aligner
- We train French-to-English aligner
- To combine the alignments, we take the intersection of the two alignments: precise: improves precision
- We separately compute the union of the two alignments: noisy: improves recall
- We build a classifier to select words from the union, which incrementally add back in the intersective alignment
- => **symmetrizing**: allows to get an alignment that maps phrases

[Och & Ney Comp. Ling. 2003]

- Unsupervised !

---

- Many improvements and variations on IBM Model 1
  - Models that take into account the position of input and output words (IBM Model 2)
  - Fertility based models (IBM Models 3,4,5): N:1: target word is aligned to N words in the source (by insertion of duplicated words)
    - + additional constraints can be modeled probabilistically: certain parts-of-speech word classes that can be switched in target
- If many different possibilities: large training data and approximate inference

# Improving the generative models

---

- Incorporation of knowledge about the structure:
  - Sequence information: HMM [Vogel et al. COLING 1996]
  - Inclusion of syntactic rules (reordering, inserting tree nodes) [Yamada & Knight ACL 2001]
- Models that enforce agreement during training [Liang et al. HLT 2006]
- Ensemble methods (combining linkers – voting) [Wu & Wang IJNLP 2005]

# Examples of learning an “interlingua”

---

- Induction of a bilingual lexicon from monolingual sources via latent concepts [Haghigi et al. ACL 2008]:
  - Maximum likelihood estimation via canonical correlation analysis (MCCA)
  - Explains matched word pairs in a common latent space
  - Training via an EM style algorithm

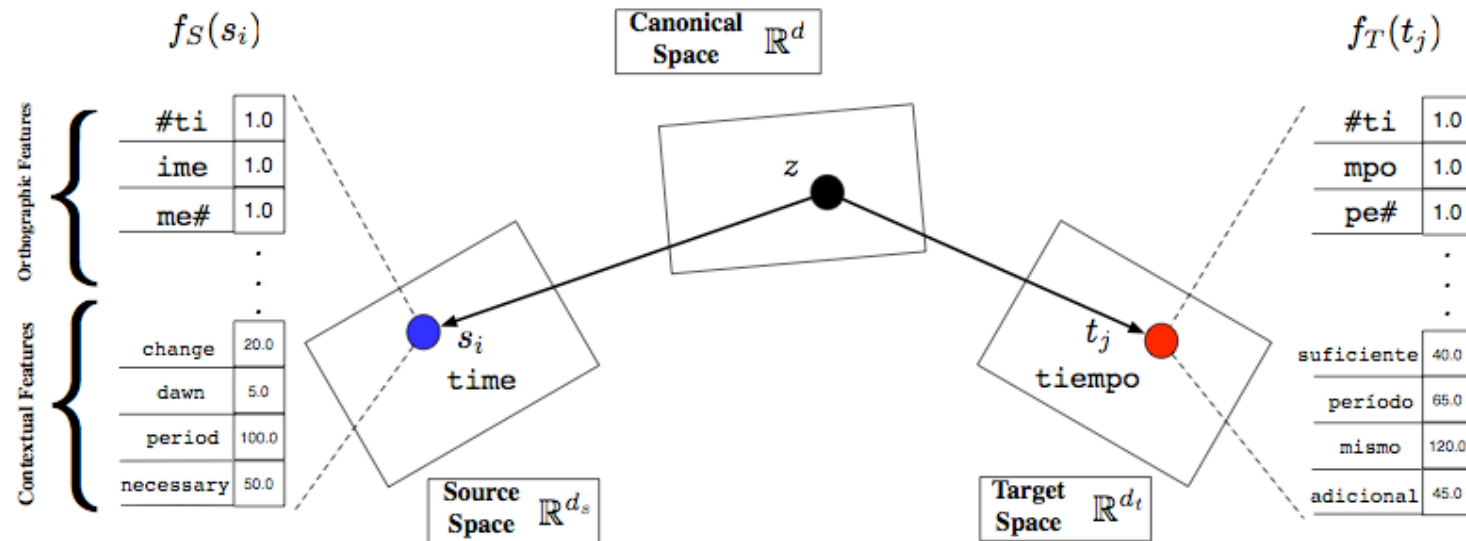


Figure 2: Illustration of our MCCA model. Each latent concept  $z_{i,j}$  originates in the canonical space. The observed word vectors in the source and target spaces are generated independently given this concept.

[Haghighi et al. ACL 2008]]

# Examples of learning an “interlingua”

- **Cept** = central pivot through which a subset of *e*-words is aligned to a subset of *f*-words
- Induction of bilingual phrase lexicon from parallel corpus based on hidden “cepts”: M:N word alignments

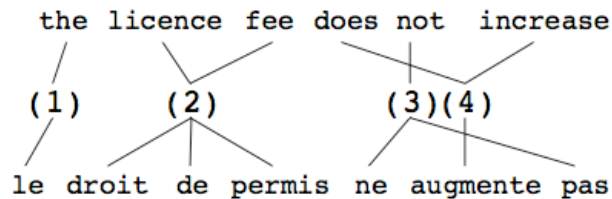


Figure 2: Same as figure 1, using cepts (1)-(4).

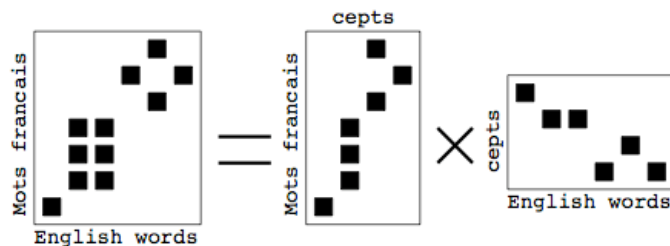


Figure 3: Matrix factorisation of the example from fig. 1, 2. Black squares represent alignments.

By means of orthogonal non-negative matrix factorization

[Goutte, Yamada & Gaussier ACL 2004]



Method	$P_S$	$R_S$	$F_S$	$P_P$	$R_P$	$F_P$	AER
ONMF + AIC	42.88%	95.12%	59.11%	75.17%	37.20%	49.77%	18.63%
ONMF + BIC	40.17%	<b>96.01%</b>	56.65%	72.20%	<b>38.49%</b>	<b>50.21%</b>	20.78%
IBM4 intersection	56.39%	90.04%	69.35%	<b>81.14%</b>	28.90%	42.62%	<b>15.43%</b>
HLT-03 best	<b>72.54%</b>	80.61%	<b>76.36%</b>	77.56%	36.79%	49.91%	18.50%

Table 3: Performance on the 447 English-French test sentences, taking NULL alignments into account, for orthogonal non-negative matrix factorisation (ONMF) using the AIC and BIC criterion for choosing the number of cepts. HLT-03 best is Ralign.EF.1 (Mihalcea and Pedersen, 2003).

Method	no NULL alignments				with NULL alignments			
	$P_S$	$R_S$	$F_S$	AER	$P_S$	$R_S$	$F_S$	AER
ONMF + AIC	70.34%	65.54%	67.85%	32.15%	62.65%	62.10%	62.38%	37.62%
ONMF + BIC	55.88%	<b>67.70%</b>	61.23%	38.77%	51.78%	<b>64.07%</b>	57.27%	42.73%
HLT-03 best	<b>82.65%</b>	62.44%	<b>71.14%</b>	<b>28.86%</b>	<b>82.65%</b>	54.11%	<b>65.40%</b>	<b>34.60%</b>

Table 4: Performance on the 248 Romanian-English test sentences (only sure alignments), for orthogonal non-negative matrix factorisation (ONMF) using the AIC and BIC criterion for choosing the number of cepts. HLT-03 best is XRCE.NoIem (Mihalcea and Pedersen, 2003).

[Goutte et al. ACL 2007]

# Discriminative models

---

- When some training data are available: discriminative models build the posterior probability directly
- **Maximum entropy** serves as a suitable framework: multinomial logistic regression, conditional random fields
  - More effective with sparse data
  - Can more easily conditioned jointly on source and target
- But in many tasks **sparsity of data** remains a problem for discriminative and generative models

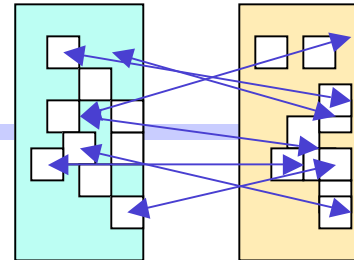
# Alignment in multilingual comparable corpora

---

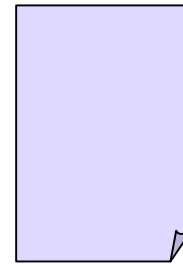
- Little research available
- Association of words:
  - Pointwise mutual information statistic
  - Chi-square
    - Possible helped by cognates, names that strongly resemble

# [Munteanu & Marcu HLT 2006]

Learning pairs  
through association  
techniques on large  
set of comparable  
documents

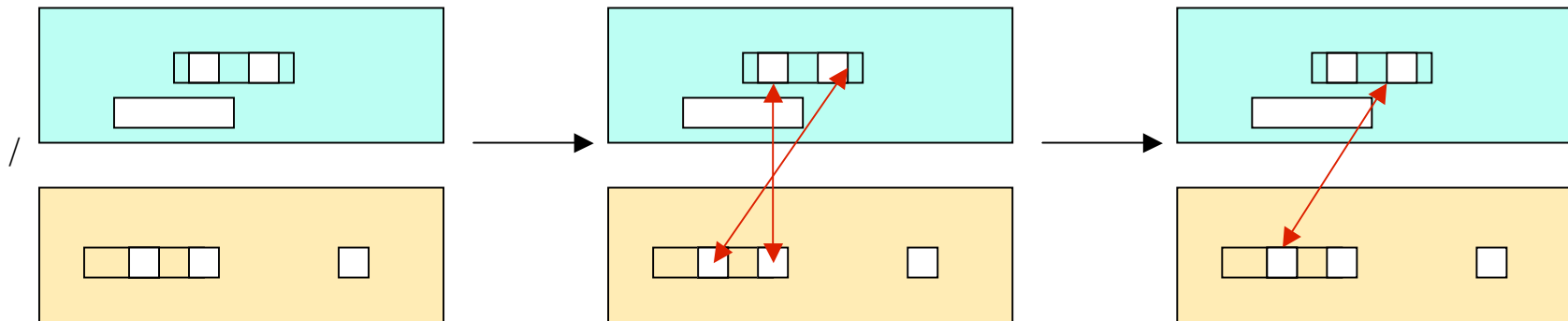


Inferring stretches  
of parallel  
fragments



Dictionary

Parallel fragments



- 
- The above approach based on latent concepts obtained through MCCA [Haghigi et al. ACL-HLT 2008] is also applied on comparable multilingual corpora
  - Cf. recent work on topic alignments in comparative multilingual corpora [De Smet & Moens SWSM 2009]

(a) Corpus Variation

Setting	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	Best- $F_1$
EN-ES-G	75.0	71.2	68.3	—	49.0
EN-ES-W	87.2	89.7	89.0	89.7	72.0
EN-ES-D	91.4	94.3	92.3	89.7	63.7
EN-ES-P	97.3	94.8	93.8	92.9	77.0

(b) Seed Lexicon Variation

Corpus	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	Best- $F_1$
EDITDIST	58.6	62.6	61.1	—	47.4
MCCA	91.4	94.3	92.3	89.7	63.7
MCCA-AUTO	91.2	90.5	91.8	77.5	61.7

(c) Language Variation

Languages	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	Best- $F_1$
EN-ES	91.4	94.3	92.3	89.7	63.7
EN-FR	94.5	89.1	88.3	78.6	61.9
EN-CH	60.1	39.3	26.8	—	30.8
EN-AR	70.0	50.0	31.1	—	33.1

- **Same Sentences:** EN-ES-P
- **Non-Parallel Similar Content:** EN-ES-W
- **Distinct Sentences, Same Domain:** EN-ES-D
- **Unrelated Corpora:** EN-ES-G

[Haghighi et al. ACL-HLT 2008]

Table 2: (a) varying type of corpora used on system performance (section 6.1), (b) using a heuristically chosen seed compared to one taken from the evaluation lexicon (section 6.2), (c) a variety of language pairs (see section 6.3).

---

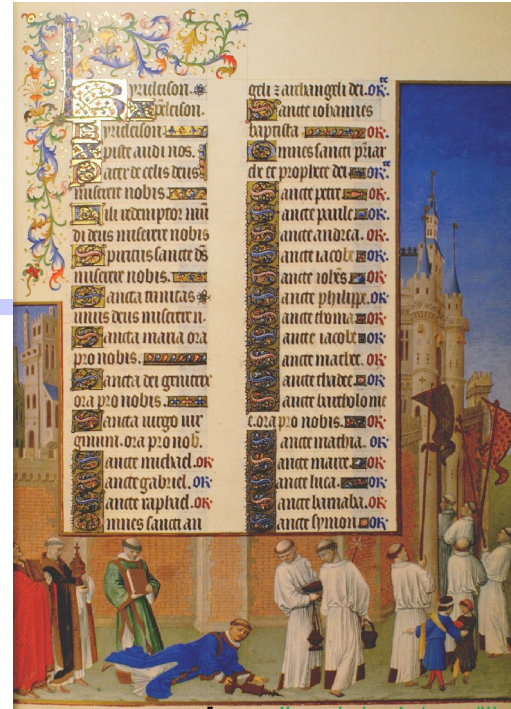
# **5. Cross-media linking of content**

- 
- Throughout history multimodality in communication is important
  - Now extra stimulated with the advent of digital media
  - “Natural language is augmented with other symbolic communication“, e.g., natural language text and images





[historylink101.net]



[www.liu.edu/cwis/cwp/library]

## Tiny horse trains as guide for blind Muslim woman

By BEN LEUBSDORF – Apr 10, 2009

DEARBORN, Mich. (AP) — Seeing-eye dogs are a nonstarter among many Muslims who consider the animals unclean, but a horse the size of a dog just might work.

"This is a really awesome little horse," Mona Ramouni said this week as she put Cali, a 3-year-old miniature horse, through her paces and rode the bus to work with her for the first time.

Ramouni lost her sight to retinopathy — damage to the retina — that is a frequent side effect of premature birth. Until now, she has relied on her family to guide her around the Detroit suburbs where she's lived, studied and worked for all of her 28 years.

Ramouni, a proofreader of textbooks in Braille, wanted more independence, but a traditional guide dog wasn't an option. She's an observant Sunni Muslim and respects her Jordanian-born parents' aversion to having a dog in the home where she lives along with three of her six siblings.

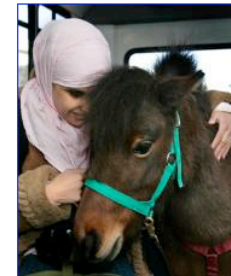
The answer, she hopes, is Cali, short for Mexicali Rose. The former show horse stands about 2 1/2 feet tall and weighs about 125 pounds.

"I want a horse that will be a partner for the next 30 or so years. ... What I really want is to be able to take her places and go places with her that neither of us ever would have been able to do without each other," Ramouni said.

While most Muslims believe dogs are unclean, some consider them "best animals."

**AP** Associated Press

Photo 1 of 5



Mona Ramouni rides a SMART bus to her job with her guide horse, Cali, in Lincoln Park, Mich., Thursday, April 9, 2009. Ramouni lost her sight soon after birth, but she can't use a guide dog. Many Muslims consider dogs unclean, and

# Today

---

- 2008: AAI fall symposium in multimedia information extraction
- 2009: Machine Learning Summer School/Workshop  
2009 University of Chicago: Workshop in Machine Learning in Computer Vision, Speech, Text and Natural Language Processing
- Workshop on Cross-Media Information Access and Mining (CIAM 2009), Twenty-first International Conference on Artificial Intelligence
- The Eleventh International Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction
- ...

# Today

---

- Emphasis on joint processing of the different modalities, similar algorithms, evaluation, ...
- One special case = aligning or linking equivalent content
- When dealing with text-image modalities:
  - Can be helpful to:
    - Automatically annotate similar images, which then can be indexed, mined, etc.
    - Summarization of multimedia: what text belongs to summarized video images and vice versa
    - Eventually build dictionary of text image pairs

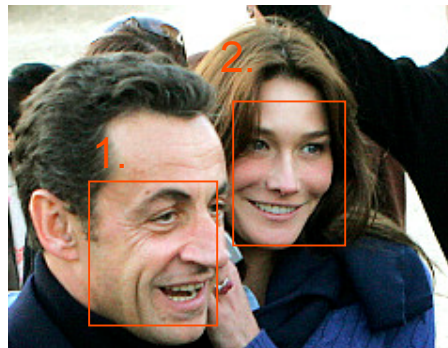


U.S. President **George W. Bush** (2nd R) speaks to the press following a meeting with the Interagency Team on Iraq at Camp David in Maryland, June 12, 2006. Pictured with **Bush** are (L-R) Vice President **Dick Cheney**, Defense Secretary **Donald Rumsfeld** and Secretary of State **Condoleezza Rice**.

[Labeled faces in the wild dataset]

# Alignment of names and faces

---



...French President 1. Nicolas Sarkozy and girlfriend 2. Carla Bruni on a trip in Egypt...





Vice President **Dick Cheney** speaks at a luncheon for Republican U.S. Senate candidate **John Cornyn** Friday, July 19, 2002, in Houston. (AP Photo/Pat Sullivan)



President-elect **Barack Obama** is inching closer to naming former rival Sen. **Hillary Clinton** as his secretary of state, ABC News has learned. (Getty Images)

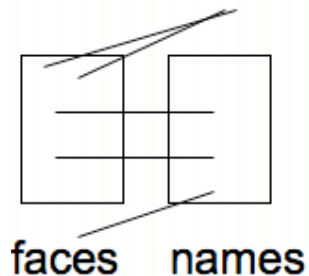


Danish director **Lars Von Trier** (C), Australian actress **Nicole Kidman** and Swedish actor **Stellan Skarsgard** (L) pose on a terrace of the Palais des festivals.(AFP/Boris Horvat)

# Alignment of names and faces

---

- Given image-text pairs  $s_i$ : align faces ( $f$ ) and names ( $n$ )
- Constraints:
  - Within image/text resolution of a face/name  $\Rightarrow$  same name and face occur once in each image/text
  - One name can only be aligned with one face, but faces can be aligned with “null” name and names with “null” face



[Pham, Moens & Tuytelaars IEEE Transactions on Multimedia 2010]

# Preprocessing

---

- **Images:**
  - **face detection**
  - **clustering of similar faces** (within and) across images (based on face descriptors)
  - computation of the **namedness** of the faces
- **Texts:**
  - **named entity (person) recognition** maximum entropy classifier augmented with gazetteers
  - **clustering of similar names** within and across texts: **noun phrase coreference resolution**
  - computation of the **picturedness** of the names

[Moens JNLE 2008] [Deschacht & Moens ACL 2007]



Cardinal from Cologne Joachim Meisner cries during a meeting with Pope Benedict XVI at the centre for dialog and prayer in Oswiecim, Poland May 28, 2006.

```
<?xml version="1.0" encoding="UTF-8"?><output><si="0">Cardinal from Cologne <ENAMEX ID="0" TYPE="PERSON">Joachim Meisner</ENAMEX> cries during a meeting with Pope <ENAMEX ID="1" TYPE="PERSON">Benedict</ENAMEX> XVI at the centre for dialog and prayer in <ENAMEX ID="2" TYPE="LOCATION">Oswiecim</ENAMEX>, <ENAMEX ID="3" TYPE="LOCATION">Poland</ENAMEX> May 28, 2006.</si>
```



[Yahoo! News]

Picturedness of name:  
Joachim Meisner: 0.75  
Benedict: 0.33



Fig. 2. An example of assigning faces to the names.

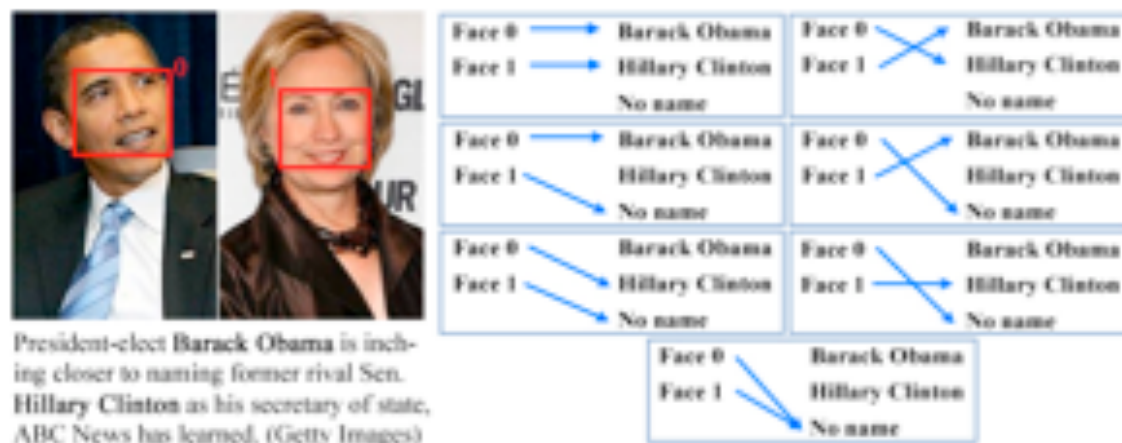


Fig. 3. An example of assigning names to the faces.

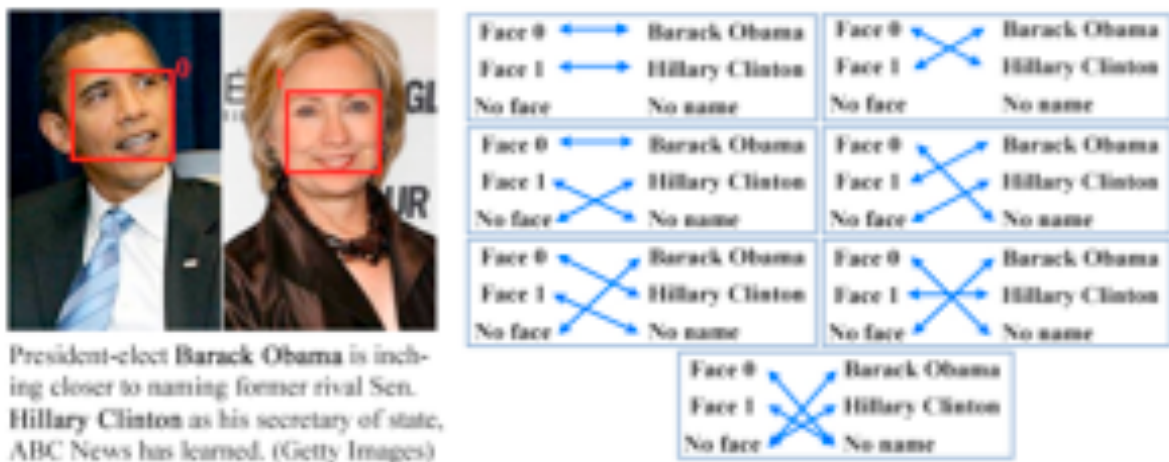


Fig. 4. An example of using evidence of name-face co-occurrence.

Likelihood of image-text pair  $s_i$  and the alignment  $a_j$ :

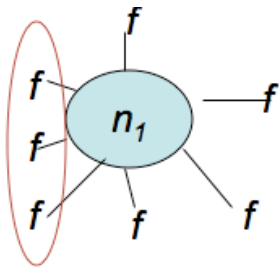
---

$$L_{s_i, a_j}^{(n \rightarrow f)} = \prod_{\alpha} P(f_{\sigma(\alpha)} | n_{\alpha})$$

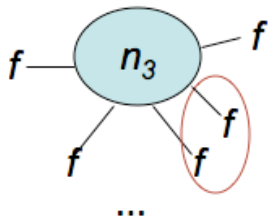
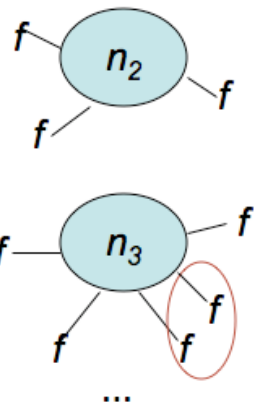
$$L_{s_i, a_j}^{(f \rightarrow n)} = \prod_{\beta} P(n_{\sigma(\beta)} | f_{\beta})$$

$$L_{s_i, a_j}^{(n^*, f^*)} = \prod_{\alpha, \sigma(\alpha) \neq NULL} [P(f_{\sigma(\alpha)} | n_{\alpha}) P(n_{\alpha}) \\ P(\text{pictured}_{\alpha} | t_{s_i}) P(\text{named}_{\sigma(\alpha)} | p_{s_i})] \\ \prod_{\alpha, \sigma(\alpha) = NULL} [(1 - P(\text{pictured}_{\alpha} | t_{s_i})) \\ P(f_{NULL} | n_{\alpha})]$$

...



e.g., estimating  $P(f|n)$



Initializing and updating the likelihood of an alignment:

- Initialization based on clustering the faces and clustering of the faces
- EM augmented with deterministic annealing
- Using all stories of the “Labeled faces in the wild” dataset: 11820 stories or image-text pairs with 5637 unique person faces and 8878 unique person names
- Unsupervised !

- 
- Use of the EM algorithm to maximize the log-likelihood of all image-text pairs  $S$ :

$$\sum_{s_i \in S} \sum_{a_j \in A_i} \delta_{i,j} \log(L_{s_i, a_j}^{(n \rightarrow f)})$$

where  $A_i$  = set of all possible alignments for image-text pair  $s_i$ ;  $\delta_{i,j}$  = strength of the alignment  $a_j$  for image-text pair  $s_i$

- The E-step updates  $\delta_{i,j}$  as follows:

$$\delta_{i,j} = \frac{L_{s_i, a_j}^{(n \rightarrow f)}}{\sum_{a_l \in A_i} L_{s_i, a_l}^{(n \rightarrow f)}}$$

- 
- During the M-step the parameter  $P(f|n)$  is recomputed:

$$P(f|n) = \frac{\sum_{s_i \in S} \sum_{a_j \in A_i} \delta_{i,j} c(a_j(n) = f)}{\sum_{s_i \in S} \sum_{a_j \in A_i} \delta_{i,j} c(n, a_j)}$$

where  $c(a_j(n) = f)$  is 1, if a face from the same face cluster  $f$  is assigned to a name of the same name cluster  $n$  in the link scheme  $a_j$ , otherwise it is 0;  $c(n, a_j)$  is 1, if the name  $n$  is assigned to a non-NULL face in  $a_j$ , otherwise it is 0

- 
- Each iteration of the EM represents a hill-climbing algorithm in the parameter space that locally minimizes this loss: we can use an **annealing like process** for finding a low-loss model where the E-step updates  $\delta_{i,j}$  as follow

$$\delta_{i,j} = \frac{[L_{s_i,a_j}^{(n,f)}]^\gamma}{\sum_{l \in A_i} [L_{s_i,a_l}^{(n,f)}]^\gamma}$$

In experiments below:  $\gamma$  is initialized at 0.02 and in each step  $\gamma = \gamma \times 1.02$  until  $\gamma = 1$



(a) Recall, precision and  $F_1$  measure of the evaluation including null name and null face.

Likelihood type	After initialization			After applying EM		
	P	R	F1	P	R	F1
$L^{(n \rightarrow f)}$	69.30%	66.42%	67.83%	69.03%	67.99%	68.51%
$L^{(f \rightarrow n)}$	69.29%	66.39%	67.81%	68.71%	66.54%	67.61%
$L^{(n,f)}$ using $P(f n)$	69.30%	66.42%	67.83%	69.25%	68.21%	68.72%
$L^{(n,f)}$ using $P(n f)$	69.29%	66.38%	67.80%	68.66%	66.70%	67.67%
$L^{(n^*,f)}$	68.10%	70.62%	69.34%	73.12%	68.87%	70.93%
$L^{(f^*,n)}$	67.55%	69.83%	68.67%	67.62%	69.90%	68.74%
$L^{(n^*,f^*)}$ using $P(f n)$	69.99%	72.79%	71.36%	74.90%	70.56%	72.66%
$L^{(n^*,f^*)}$ using $P(n f)$	69.77%	72.53%	71.12%	69.99%	72.73%	71.33%

(b) Recall, precision and  $F_1$  measure of the evaluation excluding null name and null face.

Likelihood type	After initialization			After applying EM		
	P	R	F1	P	R	F1
$L^{(n \rightarrow f)}$	65.66%	70.64%	68.06%	68.21%	69.86%	69.03%
$L^{(f \rightarrow n)}$	65.62%	70.64%	68.03%	66.08%	69.82%	67.89%
$L^{(n,f)}$ using $P(f n)$	65.66%	70.64%	68.06%	68.55%	70.21%	69.37%
$L^{(n,f)}$ using $P(n f)$	65.61%	70.63%	68.02%	66.39%	69.74%	68.02%
$L^{(n^*,f)}$	72.75%	67.18%	69.86%	66.81%	74.01%	70.22%
$L^{(f^*,n)}$	72.54%	67.43%	69.89%	72.55%	67.43%	69.89%
$L^{(n^*,f^*)}$ using $P(f n)$	75.59%	69.36%	72.34%	68.72%	<b>76.12%</b>	72.23%
$L^{(n^*,f^*)}$ using $P(n f)$	75.24%	69.09%	72.04%	75.52%	69.41%	<b>72.33%</b>

TABLE VII

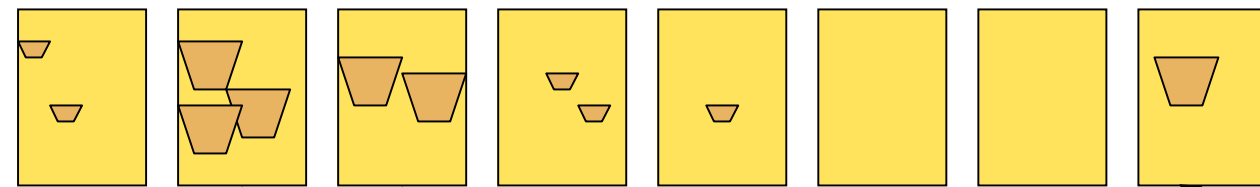
RECALL, PRECISION AND  $F_1$  MEASURE OF THE NAME-FACE ALIGNMENT WHERE THE EM IS AUGMENTED WITH DETERMINISTIC ANNEALING IN THE LABELED FACES IN THE WILD DATASET;  $\gamma = 0.02 \rightarrow 1.0$ ; AT EACH  $\beta$  VALUE.  $n^*$  DENOTES THE USE OF PICTUREDNESS VALUE IN THE LIKELIHOOD FUNCTIONS AND  $f^*$  DENOTES THE USE OF NAMEDNESS VALUE IN THE LIKELIHOOD FUNCTIONS.

[Pham, Moens & Tuytelaars IEEE Transactions on Multimedia 2010]

# Cross-media alignment of names and faces

## ■ Adaptation to alignment in **video**

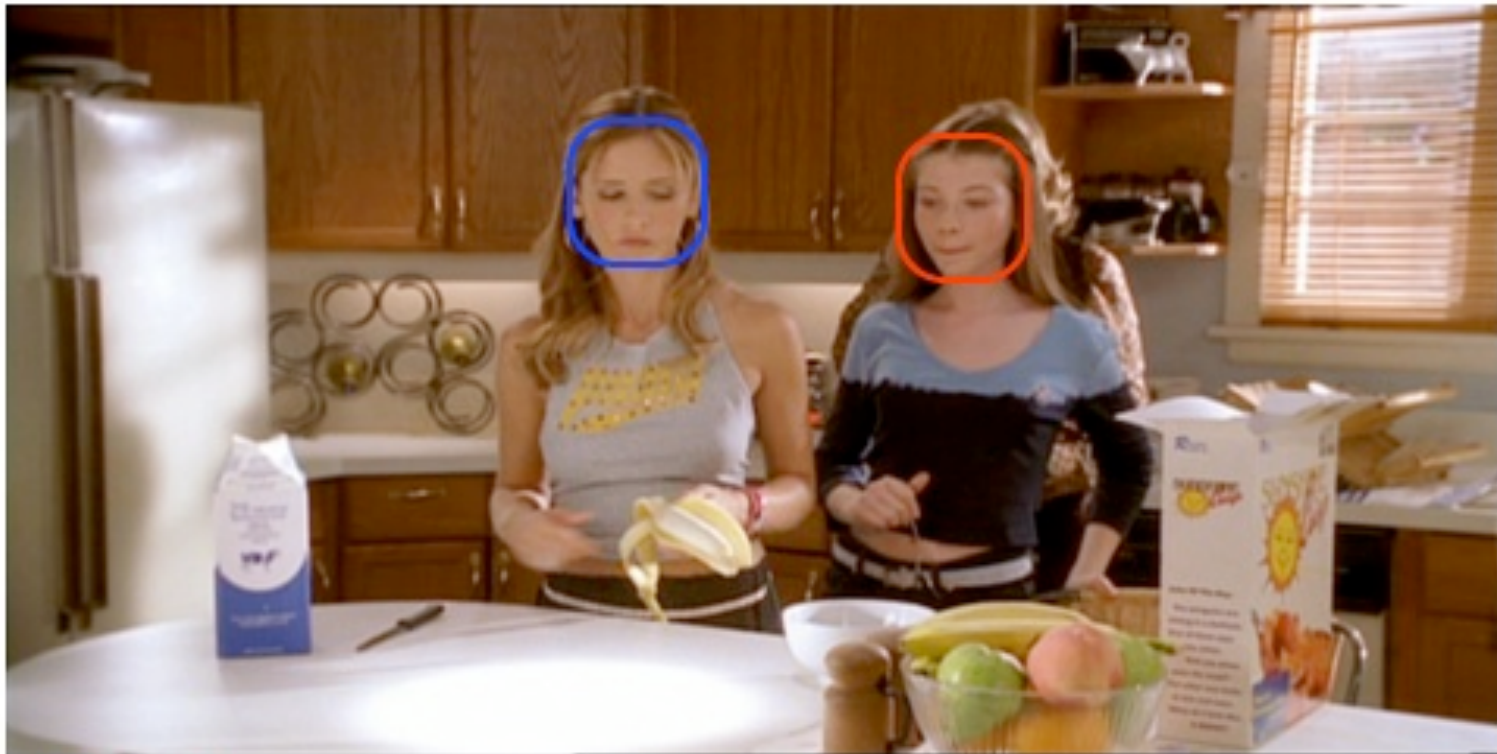
Time warping  
to align scripts  
and subtitles



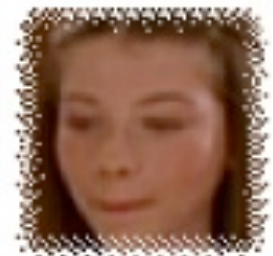
**Dawn** returns with a spoon, wearing an innocent expression.  
**Buffy** turns away to get another bowl.

Use of **unsupervised alignment** over all “stories” in one act  
(story = 1 frame + text):

some faces receive null name => use of face **label propagation**  
with random walk algorithm



**Buffy**



**Dawn**

**Dawn** returns with a spoon, wearing an innocent expression.  
**Buffy** turns away to get another bowl.

[Pham et al. 2010 publication in preparation]



00:19:35:950 Tonight, **Robert Mugabe** confirmed that the presidential election will be as planned this Friday, despite the withdrawal of the opposition leader, **Morgan Tsvangirai**.

00:19:45:910 Here is Africa correspondent, **Orla Guerin**.

00:19:56:680 Increasingly isolated but defiant as ever, **Robert Mugabe** had a message for his critics today: "Mind your own management".

[BBC News 2008]

- This framework uses the set of labeled faces  $F_l$  with name labels  $N_l$  and the set of unlabeled faces  $F_u$  to predict the name labels  $N_u$  of  $F_u$ . The number of distinct names is known.
- A fully connected graph  $G$  is built where the nodes are all labeled and unlabeled faces. The weight  $w_{ij}$  of the edge between faces  $f_i$  and  $f_j$  is the similarity between them + taking into account extra constraints.
- The one-step transition probability  $T_{ij}$  from face  $f_i$  to face  $f_j$  can be estimated from the edge weights:

$$T_{ij} = P(j \rightarrow i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}}$$

- 
- We define a probability matrix  $N$  of size  $(|F_l|+|F_u|) \times K$  where  $N_i$  is the probability distribution of name labels over face  $f_i$ .
  - The label propagation algorithm:
    1. All faces/name-face pairs propagate labels for one step:  $N \leftarrow TN$
    2. Row normalize  $N$  to maintain the label probability interpretation
    3. Clamp the labeled faces. Repeat from step 2 until convergence of  $N$
  - After the label propagation, the matrix  $N$  contains the label distribution for each face
  - We use the name with highest probability where the probability is above a threshold  $\lambda$  (“refusal to predict” mechanism).

- 
- ▶ We perform and compare our experiments on two BBC news broadcasts recorded on 22-Jun-2008 and 27-Jun-2008.
  - ▶ Our best face labeling results obtained at 100% “refusal to predict” recall (all test examples are labeled with most probable label):
    - ▶ [BBC\\_22-Jun-2008](#) broadcast: precision of 82.56%
    - ▶ [BBC\\_27-Jun-2008](#) broadcast: precision of 51.09%
  - ▶ Cf. SVM trained on the same manually labeled faces:
    - ▶ [BBC\\_22-Jun-2008](#) broadcast: precision of 55.81%
    - ▶ [BBC\\_27-Jun-2008](#) broadcast: precision of 26.09%

---

# 6. Conclusions



- 
- Linking content = important intelligent task for a machine
    - Focus on equivalence relation
    - Examples in monolingual, cross-lingual, cross-media settings, but many other applications
    - Generic, flexible underlying algorithms, adapted to the specific setting of the task that often require little supervision

- 
- Still many research questions:
    - How to deal with sparsity of the data, with efficiency, N:M relations, ...?
    - How to discover other “discourse” relations in the unstructured sources? Could also be discovered from the data, but sparsity is even bigger problem.
    - How to combine extractions/recognitions with the linking? To combine with metadata, descriptors?
    - ...

**MANY RESEARCH VENUES TO EXPLORE ...**



- 
- Thanks to CLASS, AMASS++, DAISY, TermWise, WebInsight, PuppyIR and TERENCE projects and the researchers involved (Phi The Pham, Ivan Vulic, Wim De Smet, Karl Gyllstrom and Koen Deschacht) and colleague Tinne Tuytelaars



Daisy

