An evaluation model for systems and resources involved in the correction of errors in textual documents

Arnaud Renard, Sylvie Calabretto, Béatrice Rumpler

Laboratoire d'InfoRmatique en Image et Systèmes d'information

LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon
Université Claude Bernard Lyon 1, bâtiment Nautibus
43, boulevard du 11 novembre 1918 — F-69622 Villeurbanne cedex
http://liris.cnrs.fr

Context & Issues

- **Information production process evolution leads to errors**

- **How to manage errors in IR systems ?**

  - **Errors in queries:**
    - → Suggestions of more likely queries through « Did you mean… »

  - **Errors in documents:**
    - → Queries expansion
      - *Expand queries with keywords variations standing for errors expected to lie in documents*
    - → Documents error processing
      - *Correction of errors directly in processed documents*
      - *Seems to be the best fitted one according to [Kantor'00]*

- **How to evaluate error correction systems ?**

Outline

State of the art
- Types of errors & classification
- Error correction approaches
- Error correction evaluation limits

Proposal
- Generic Evaluation Model (Meta-Model)
- Specific Evaluation Model (Model)

Evaluation
- Evaluation model implementation
- Instantiation & analysis of evaluation model resources
- Results

Conclusion & further works

Types of errors & classification (1/2)

Non-word:

- **Invalid word according to a lexicon.**
- **Example:**

  « The bok is on the table. »

  - The word « *bok* » doesn't exist in English and probably comes from mistyping the word « *bo~~o~~k* ».
  - The sentence should be: « *The book is on the table.* ».

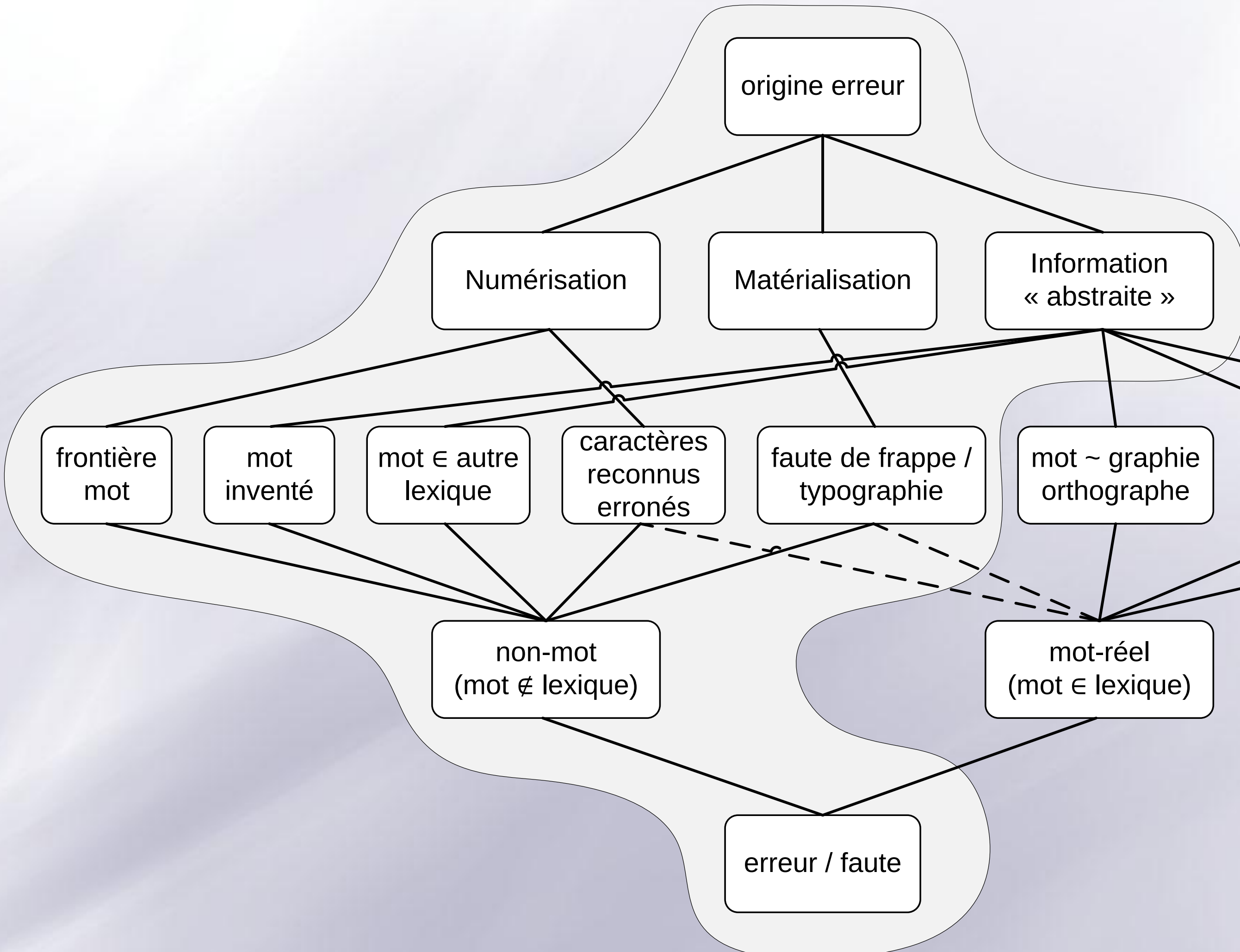
Real-word:

- **Valid word according to a lexicon but not the intended word.**
- **Example:**

  « I saw tree trees in the park. »

  - The word « *tree* » exists in English but doesn't mean anything in this context so it probably comes from mistyping the word « *t~~h~~ree* ».
  - The sentence should be: « *I saw three trees in the park.* ».

Types of errors & classification (2/2)



- **Erreurs mots-réels → plutôt liées à l'information « abstraite »**
- **Erreurs non-mots → plutôt liées à la matérialisation / numérisation de l'information**

Error correction approaches

Semantic based approaches

- **Intuition [Hirst'98, Hirst'05]:**
  - Intended words are generally semantically related to close words which constitute the context
  - Real-word errors brake the semantics
  - Application of semantic disambiguation methods to correct errors
- **Constraints:** context + needs semantic resource

Statistic based approaches

- **Without context [Mitton'09]: simple word frequency**
  - 84% of correct words belongs to most frequent words [Pedler'07]
  - 62% of errors belongs to less frequent words [Pedler'07]
- **With context: statistics on words co-occurence (n-grams)**
  - Trigram presence in a large corpus of documents (BNC) [Verberne'02]
  - Trigram probabilities [Mays'91]
  - N-grams probabilities [Golding'99], [Islam'09] trained over Google Web 1-T n-grams)
- **Constraints:** context + learning phase

Error correction evaluation limits

Different approaches are difficult to compare each other [Wilcox'08]

- **Different resources involved**
  - Reference dictionaries
  - Error collections (randomly generated errors in a pre-existing collection of documents)
    - *Is it representative ?*
    - *Is it reproducible ?*
  - Evaluation metrics
  - Autonomous error correction systems (black boxes services)

- **Elements to address this issue**
  - Gather and distribute collection of real documents which contains errors [Pedler, 2007]
  - Implement previous approaches at the same time to have similar experiment conditions [Wilcox, 2008]

→ **Standard evaluation model to use to evaluate error corrections systems in the same way.**

Proposal (1/4): Evaluation model

Why:

- **Formalize an evaluation framework**
  - Inspired by Cranfield [Cloverdon'60, Cloverdon'66] → TREC, INEX, …
- **Evaluation:**
  - **composites systems**: open systems created from an original resources combination
  - **autonomous systems**: closed systems which can be seen as black boxes

3 levels:

- **Meta-model (Generic Evaluation Model GEM)**
  - Defines different resources types
- **Model adapted to evaluate error correction systems (Specific Evaluation Model SEM)**
  - Derived from GEM
  - Defines families of resources for each type
- **Instantiation of model resources**

Proposal (2/4) : GEM (Meta-model)

Generic Evaluation Model:

$$GEM = \langle R_D, R_P, s, R_E, a \rangle$$

- **R_D: data resources**
  - Example : data to process
- **R_P: processing resources**
  - Example: algorithms to apply to data
- **R_E: evaluation resources**
  - Example: evaluation metrics, reference values
- **s: data processing module based on provided resources R to produce results**
  - Example: scores
- **a: module to evaluate data processing s results and produce performance indicators**
  - Example: recall, precision, MRR, …

Proposal (3/4): SEM (Model)

Specific Evaluation Model:

- **Composites Systems evaluation:**

$$SEM_{composite} = \langle \{Coll, Dict\}, SDM, s, EM, a \rangle$$

- **R_D: data resources**
  - $Coll$: Error collection (list of pairs of the form: $\langle wrong\ word, target\ word \rangle$)
  - $Dict$: Reference dictionary (list of the form: $\langle word, word\ frequency \rangle$)

- **R_P: data processing resources**
  - SDM: Similarity and Distance Measures normalized [0, 1]
  - AS: Autonomous System

- **R_E: evaluation resources**
  - EM: Evaluation Metrics

Proposal (4/4): SEM (Model)

Specific Evaluation Model

- **Autonomous Systems evaluation:**

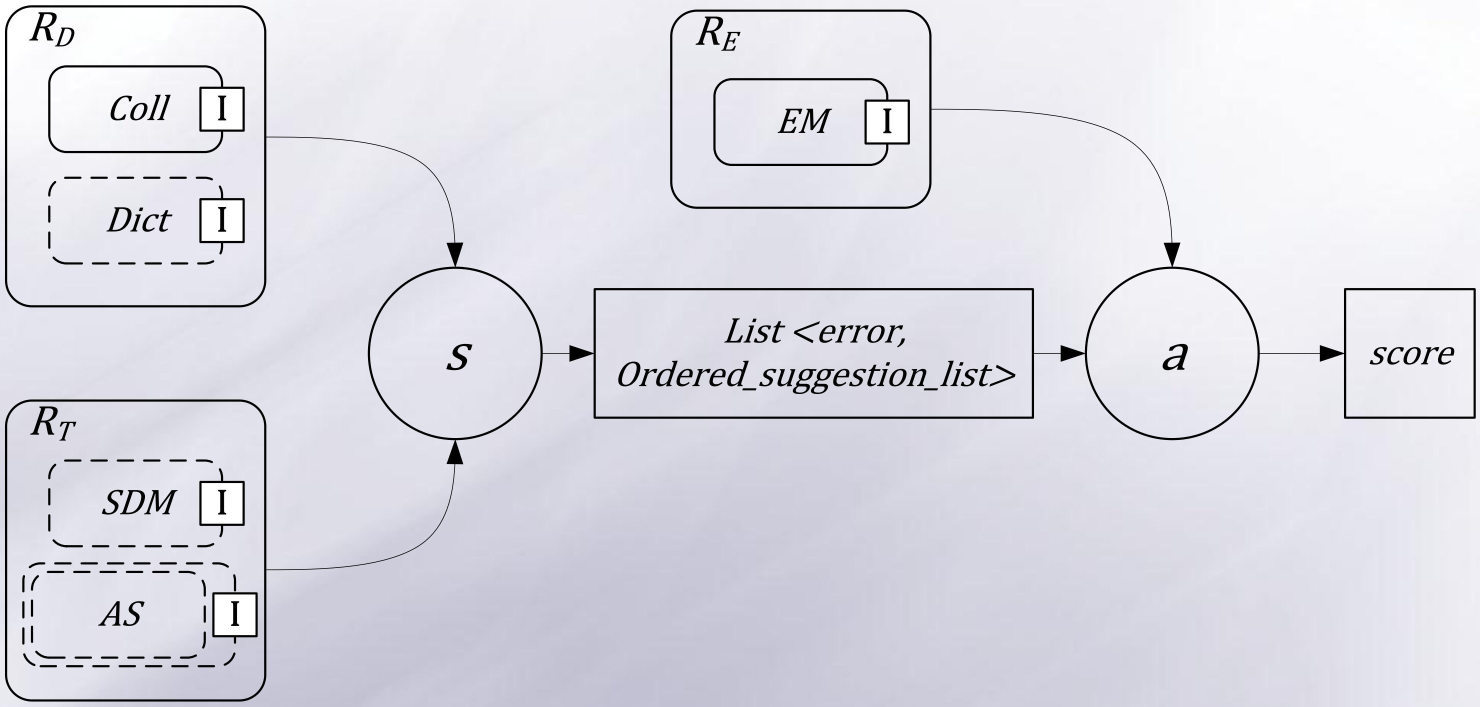
$$SEM_{autonomous} = \langle \{Coll\}, AS, s, EM, a \rangle$$

- **R_D: data resources**
  - $Coll$: Error collection (list of pairs of the form: $\langle wrong\ word, target\ word \rangle$)
  - $Dict$: Reference dictionary (list of the form: $\langle word, word\ frequency \rangle$)

- **R_P: data processing resources**
  - SDM: Similarity and Distance Measures normalized [0, 1]
  - AS: Autonomous System

- **R_E: evaluation resources**
  - EM: Evaluation Metrics

Evaluation (1) : Model implementation



- **Standard interface for every type of resources which are implemented by one or many OSGi bundles**
- **Easy to substitute one bundle with another while they are supposed to have the same type (ex: replace one similarity/distance measure with another one)**

Evaluation (2) : Resources instantiation

Errors collection:

- **WCM Wikipedia Common Misspellings [Wikipedia'10]**
  - **Built from frequent Wikipedia contributors errors**
  - **4274 couples** $< wrong\ word, target\ word >$
  - **Both non-word and real-word errors**
    - *No provided context but errors already identified*

Evaluation (3) : Resources instantiation

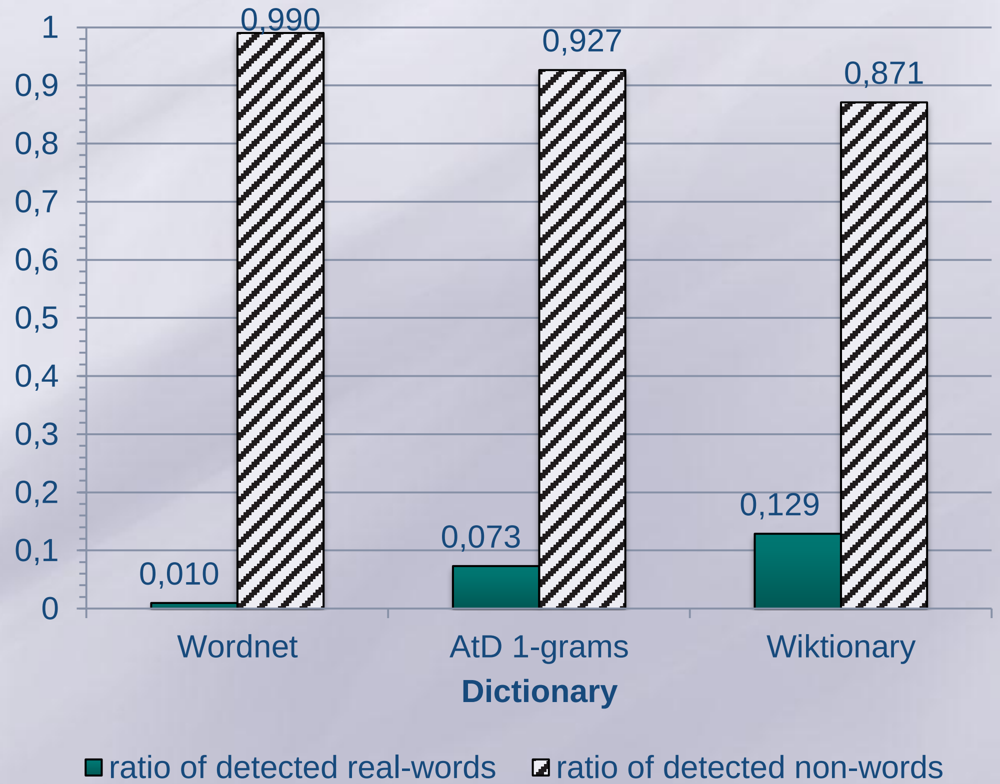
Dictionaries :

- **Wordnet [Miller'95] [Fellbaum'98]**
  - Semantic resource employed as a lexical database
  - Number of words: 147 000 words
- **AtD Unigrams [AtD'10]**
  - Learning of most frequent unigrams on a large document dataset
  - Number of words: 165 000 words
- **Wiktionary [Wiktionary'10]**
  - Online collaborative dictionary → evolves continuously
  - Number of words: 2 000 000 words

Evaluation (4) : Resources analysis

Dictionaries :

- Proportion of words in the collection which identified as real-words errors (resp. non-words) according to the dictionary.

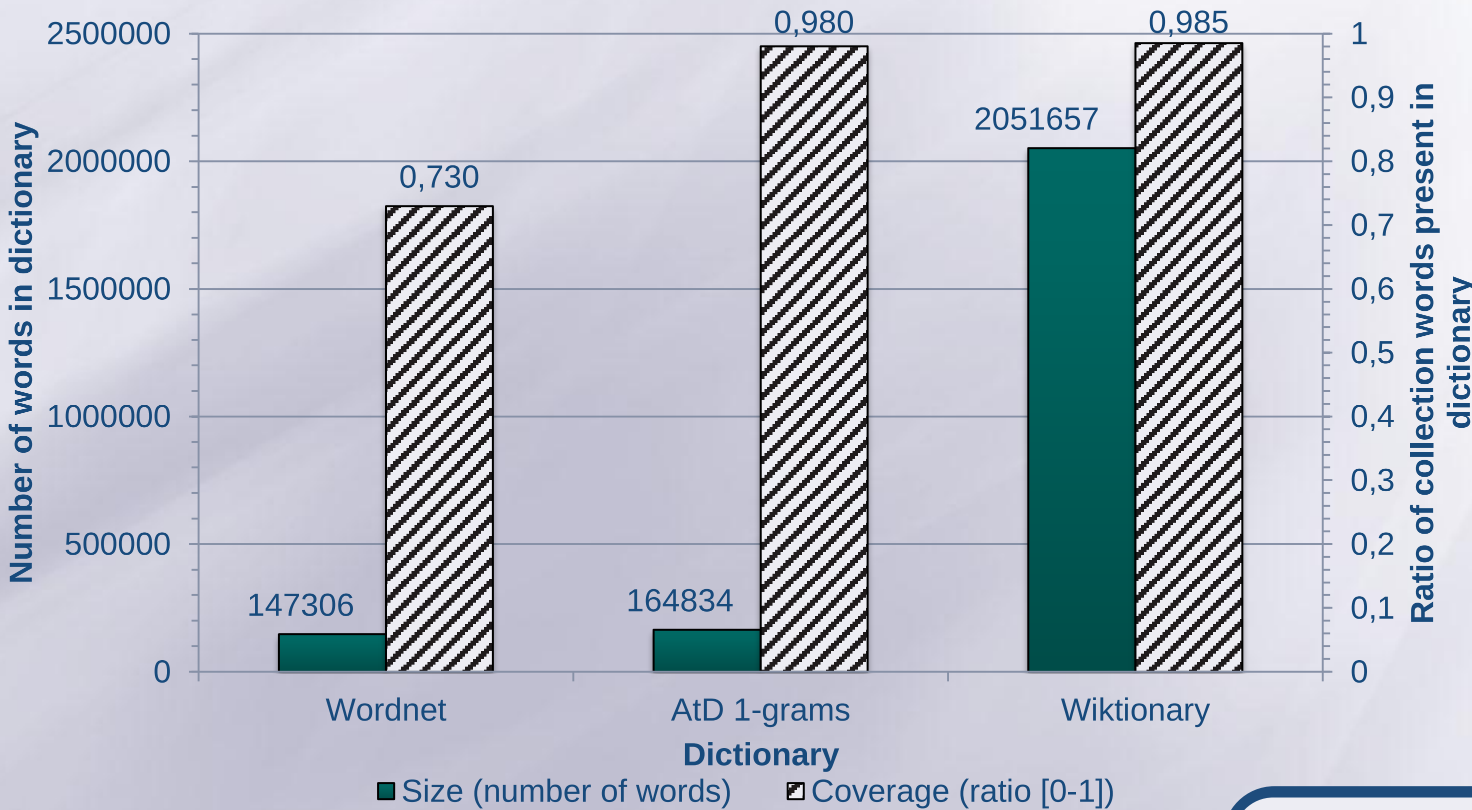


- Dependent of the dictionary
- Temporal evolution (new/old words)

Evaluation (5) : Resources analysis

Dictionaries :

- Dictionaries size and collection of errors target words coverage.



Evaluation (6) : Resources instantiation

Similarity / distance measures

- **Levenshtein distance (edit distance)**
  - Minimal number of characters to delete, insert or substitute to transform a word in an another word.
- **Jaro distance**
  - Based on the number of matching characters between two strings.
  - Two characters can be considered as matching if their positions in the words are not too far from each other (maximum distance threshold).
- **Jaro-Winkler distance**
  - Same as Jaro distance but strings having similar prefixes are favoured.

Evaluation (7) : Resources instantiation

Evaluation metric

- **Integrate an error correction system to an IR system**
- **Constraint: same as offline error correction systems**
  →No user interactions

	Online error correction (standard)	Offline error correction (ITEC)
Contextual data	Yes: directly usable	No: metadata → assumptions
User interactions	Yes: choice among many suggestions (≈ 5)	No: no choice → high precision required

Evaluation (8) : Resources Instantiation

Evaluation metric

- **Precision oriented error correction system**
  - Correct result in first position

    → Appropriate metric: $MRR\ (Mean\ Reciprocal\ Rank)$

$$MRR = \frac{1}{|errorsCouples|} \sum_{i=1}^{|errorsCouple|} \frac{1}{rank_{FoundTagetWord}}$$

  - MRR applies a huge penalty when the good result is not ranked first (divide by the rank)

Evaluation (9) : Synthesis

Instantiation (EMI) of SEM's resources

$$EMI_1 = \langle \{WCM, Wiktionary\}, Jaro - Winkler, s, MRR, e \rangle$$

$$EMI_2 = \langle \{WCM, Wiktionary\}, Jaro, s, MRR, e \rangle$$

$$EMI_3 = \langle \{WCM, Wiktionary\}, Levenshtein, s, MRR, e \rangle$$

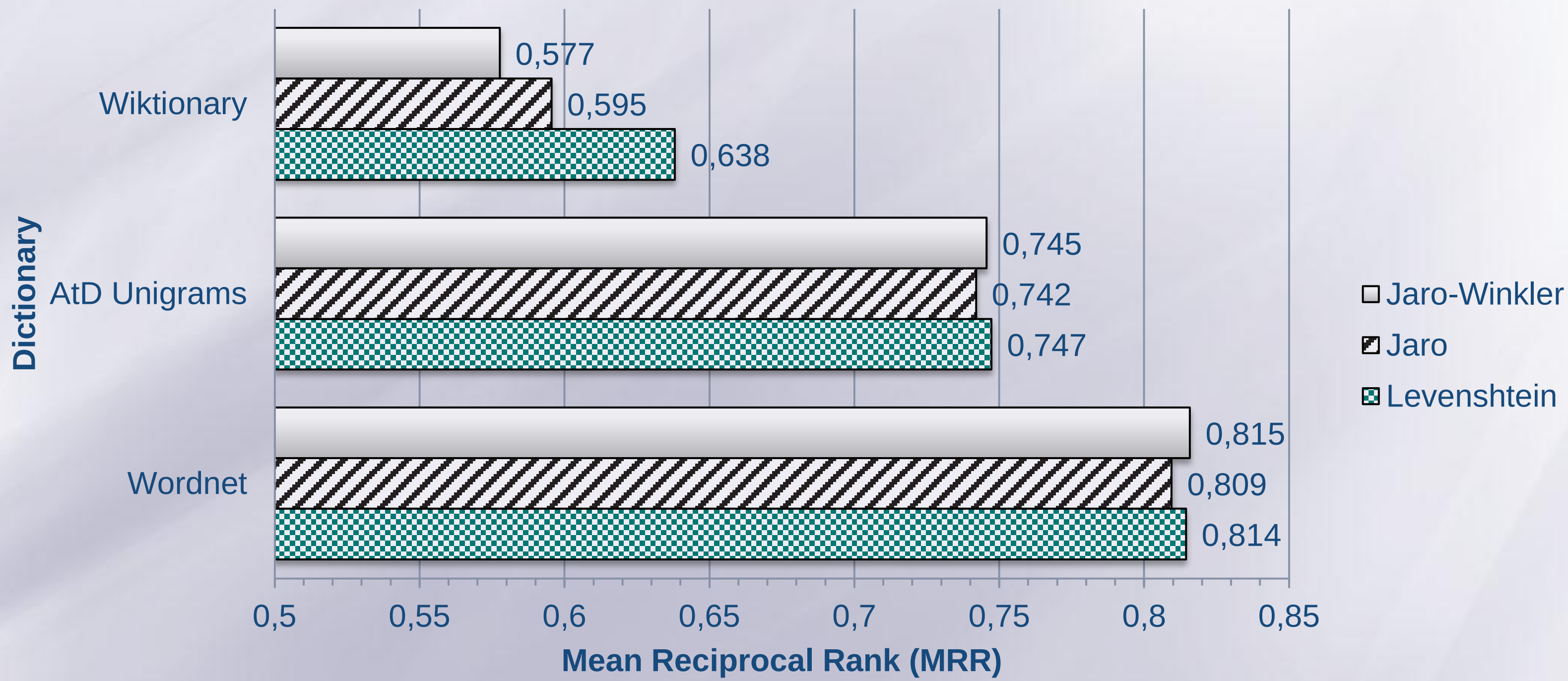
$$EMI_4 = \langle \{WCM, AtD\}, Jaro - Winkler, s, MRR, e \rangle$$

...

$$EMI_9 = \langle \{WCM, Wordnet\}, Levenshtein, s, MRR, e \rangle$$

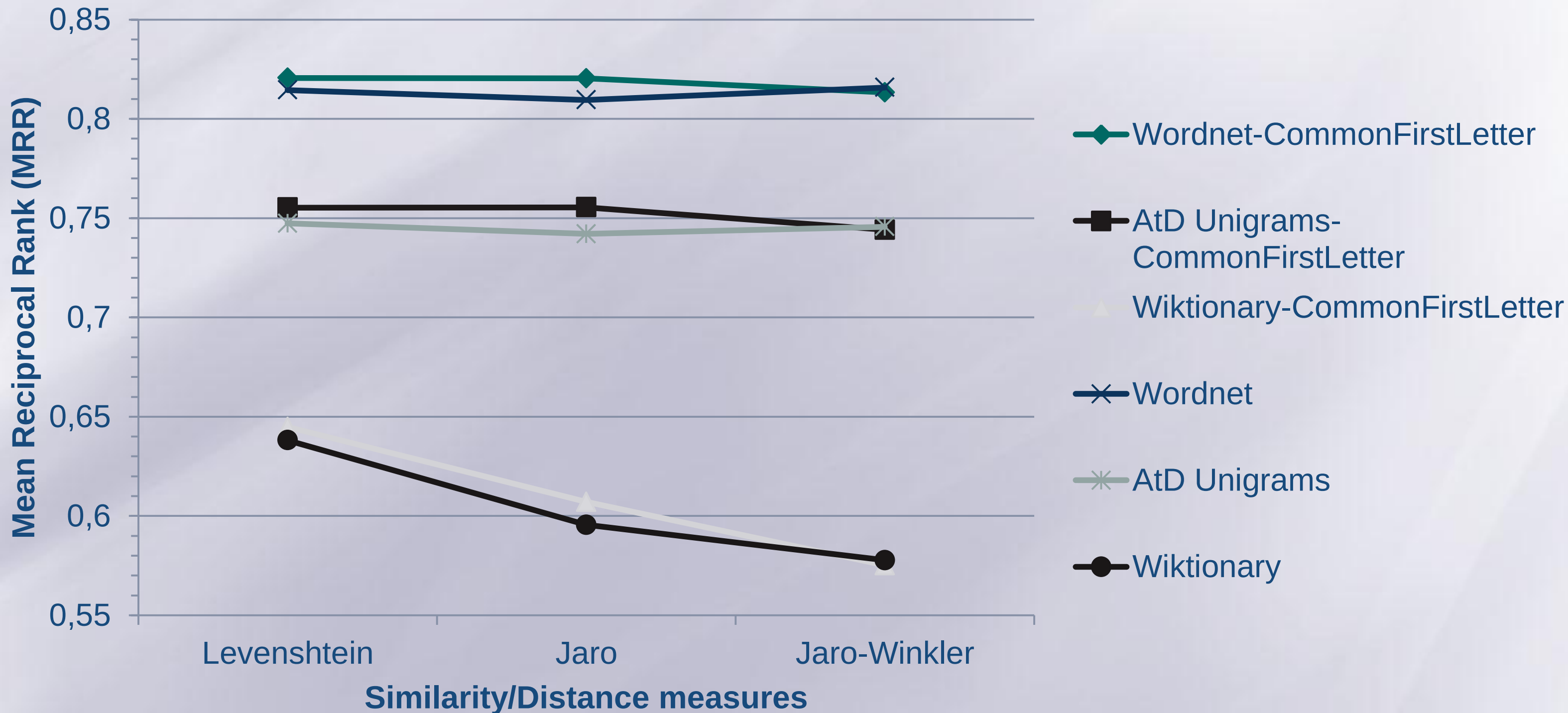
Evaluation (10) : Results

- **MRR of different combinations between similarity/distance measures and dictionaries**



Evaluation (11) : Results

- **Error collection analysis shows that 97% of errors share their first character with their intended target word**
  - → 9 new EMI integrating this heuristic
- **Comparison of MRR values with/without considering the first letter as being common**



Conclusion & further works

Errors correction systems:

→ Identification of difficulties to evaluate approaches

→ Proposal of an evaluation model

→ Implementation of this model in an extensible platform

→ First evaluations of composites error correction systems

Further works:

- Evaluate autonomous error correction systems like Google, Yahoo, Aspell, University of Western Australia prototype, AftertheDeadline (ongoing work)
- Integrate other kinds of similarity measures (like phonetic similarity measures)
  - When considering the Web errors in written content tends to be the same as spoken content [Baron, 2003]
- Evaluations based on other error collections providing additional context [Pedler, 2007]
- Integrate a composite system to an IR system to evaluate indirectly error correction systems over well known IR campaigns like INEX and TREC

Questions

Références (1)

1. De Rosnay J. La révolte du pronétariat : Des mass média aux média des masses. 2006.
2. Subramaniam L.V., Roy S., Faruquie T.A., Negi S. A Survey of Types of Text Noise and Techniques to Handle Noisy Text. Language, 115-122, 2009.
3. Ruch, P. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 7, 2002.
4. Kantor P.B., Voorhees E.M. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. Information Retrieval 2, 2, 165–176, 2000.
5. Baron, N.S. Language of the Internet. (2003), 1-63.
6. Hirst G., St-Onge D. WordNet: An electronic lexical database, chapter Lexical chains as representations of context for the detection and correction of malapropisms, pages 305–332. The MIT Press, Cambridge, MA, 1998.
7. Hirst G., Budanitsky A. Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering, 11(1):87–111, March 2005.
8. Mitton, Roger. Ordering the suggestions of a spellchecker without using context. London: Birkbeck ePrints. Available at: http://eprints.bbk.ac.uk/782, 2009.

Références (2)

9. Pedler, J. Computer Correction of Real-word Spelling Errors in Dyslexic Text. 239. http://www.dcs.bbk.ac.uk/~jenny/Publications/PedlerPhD.pdf, 2007.

10. Mays E., Damerau F. J., Mercer R. L. Context based spelling correction. Information Processing and Management, 27(5):517–522, 1991.

11. Golding A. R., Roth D. A winnow-based approach to context-sensitive spelling correction. Machine Learning, 34(1-3):107–130, 1999.

12. Islam A., Inkpen D. "Real-Word Spelling Correction using Google Web 1T n-gram Data Set," site.uottawa.ca, pp. 1689-1692, 2009.

13. Verberne S. Context-sensitive spell checking based on word trigram probabilities. Master's thesis, University of Nijmegen, February-August 2002.

14. Wilcox-O'Hearn L. A., Hirst G., Budanitsky A. Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In A. Gelbukh, editor, In Proceedings of CICLing (LNCS 4919, Springer-Verlag), pages 605–616, Haifa, 2008.

15. Wikipedia. Wikipedia List of Common Misspellings. Last accessed 15 October 2010. http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings.

16. Miller George A. WordNet: A Lexical Database for English. Communications of the ACM, vol. 38, no. 11: 39-41, 1995.

Références (3)

17. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

18. AtD. After the Deadline Unigrams Corpus. Last accessed 15 October 2010. http://static.afterthedeadline.com/download/unigrams.txt.gz.

19. Wiktionary. Wiktionary Online Collaborative Dictionary. Last accessed 15 October 2010. http://en.wiktionary.org/wiki/Wiktionary:Main_Page.

Contexte : Des erreurs ?... quel intérêt ?

Evolution des processus de production de l'information

- **Qui ?**
  - Professionnels de l'information → Utilisateurs lambda [De Rosnay, 2006]
- **Quoi ?**
  - Contenu textuel proposé par les outils et services fournis sur le Web : pages Web, Blogs, Wiki, …
- **Comment ?**
  - Cadre professionnel → Cadre privé
  - Contrôle de qualité de l'information → libre autopublication (Blogs, Wiki, …)
- **Constat :**
  - **Sources d'information plus nombreuses et plus diversifiées**
  - **Qualité de l'information inégale**
    - *domaine mal connu, vocabulaire non-maitrisé et/ou employé de façon inadaptée, pas de contrôle, pas ou peu de « correction », …*
  - → **+ d'erreurs [Subramaniam, 2009]**

→ **Impact sur les systèmes devant gérer les informations**

→ **En particulier sur la qualité des index produits (et donc sur les performances) :**

- → Accroissement de la taille des index [Ruch, 2002]
- → Silences à l'interrogation

Principaux concepts (1) : Définitions de base

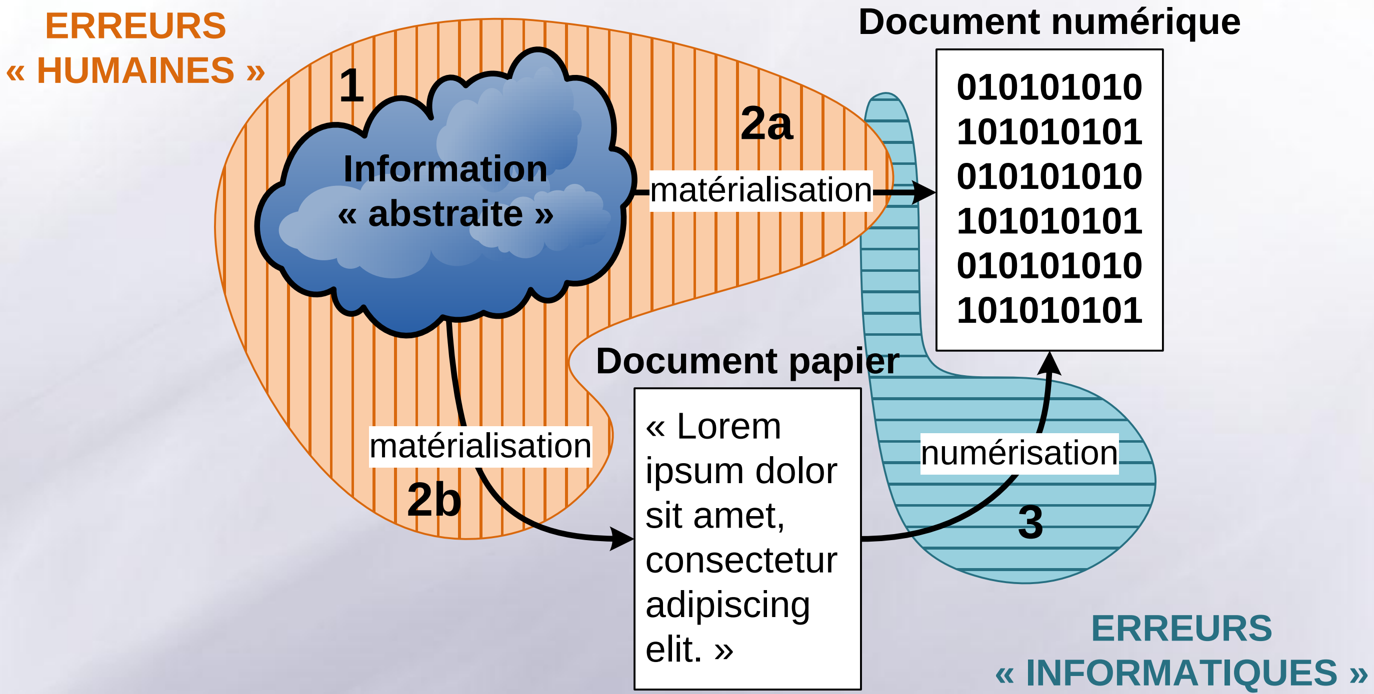
- **Alphabet *A* :**
  - **Ensemble fini des lettres *l* d'une langue.**

- **Mot *m* :**
  - **Séquence ordonnée de k lettres de l'alphabet prise parmi l'ensemble des mots de k lettres A^k.**
    - $\forall l_i \in \mathrm{A}, m \in \mathrm{A}^k \; ssi \; m = l_1, l_2, \ldots, l_{k-1}, l_k$
      - *Exemple : « tree »*
  - **Un mot est appelé mot valide s'il fait partie des mots usités de la langue**

- **Dictionnaire *d* (ou lexique) :**
  - **Ensemble des mots valides d'une langue.**

- **Erreur *e* (ou mot erroné) :**
  - **Mot dont au moins une lettre diffère de la lettre attendue à une position donnée dans la séquence des lettres correspondant à un mot cible correct**
    - $\forall l_i \in \mathrm{A}, \exists l_j \in \mathrm{A} \text{ tel que } \mathrm{i} = \mathrm{j} \text{ et } l_i \neq l_j$
      - *Exemple : « book » → » bok » (non-mot)*
      - *Exemple : « tree » → « three » (mot-réel)*

Origines (1) : Matérialisation de l'information



- **Erreurs humaines à la création (confusion sur le sens, verbalisation de l'idée, association à un mot) de l'information (1) : enfants, étrangers, dyslexiques, …**
- **Erreurs humaines à l'expression (dysorthographie, mauvaise prononciation) et à la saisie :**
  - Informatique : typographie (mauvaise touche, touche défaillante, …) (2a).
  - Manuscrite : dysgraphie (lettres manquantes, lettres mal formées, …) (2b).
- **Erreurs machine à la numérisation (OCR / ASR) de l'information (3).**

➔ Erreurs non mutuellement exclusives → cumul possible

Approches de correction d’erreurs (1)

Approches sémantiques

- **Contraintes :**
  - Contexte obligatoire
  - Utilisation d’une ressource sémantique
- **Intuition [Hirst 1998, Hirst 2005] :**
  - Les mots que le rédacteur a l’intention d’écrire sont généralement sémantiquement liés aux mots présents dans le contexte environnant
  - Erreur de type mot-réel → perte du lien sémantique
  - Application des méthodes de désambiguïsation sémantique à la correction d’erreurs

Approches de correction d’erreurs (2)

Approches statistiques

- **Contrainte :**
  - Apprentissage nécessaire

- **Sans contexte [Mitton, 2009]**
  - **Simple fréquence des mots**
    - *84% des mots corrects sont parmi les mots les plus fréquents [Pedler, 2007]*
    - *62% des erreurs sont parmi les mots les moins fréquents [Pedler, 2007]*

- **Avec contexte**
  - **Statistiques sur la cooccurrence des mots (n-grams)**
    - *Probabilités de trigrammes [Mays, 1991]*
    - *Probabilités de n-grams [Golding, 1999], [Islam, 2009] (entrainés sur le Google Web 1-T n-grams)*
    - *Existence du trigramme dans un corpus de grande taille BNC [Verberne, 2002] → pas de localisation précise de l’erreur*

→ Approches avec contexte donnent de meilleurs résultats

Evaluation (6) : Instanciation/analyse des res.

Mesures de Similarité et de Distance

- **Distance de Levenshtein (distance d'édition classique)**
  - Nombre minimal de caractères qu'il faut effacer, insérer ou substituer pour passer d'un mot à un autre

$$\mathrm{D}_{\mathrm{Levenshtein}}(x_i, y_j) = min\begin{pmatrix} \mathrm{D}(x_{i-1}, y_j) + 1 \\ \mathrm{D}(x_i, y_{j-1}) + 1 \\ \mathrm{D}(x_{i-1}, y_{j-1}) + cout \end{pmatrix}, et\ cout\ vaut\ \begin{cases} 0, si\ x_i = y_j \\ 1, sinon \end{cases}$$

	d	a	i	r	y
d	0	1	2	3	4
i	1	1	1	2	3
a	2	1	2	2	3
r	3	2	2	2	3
y	4	3	3	3	2

$$\mathrm{D}_{\mathrm{Levenshtein}}("dairy", "diary") = 2$$

$$\mathrm{D}_{\mathrm{LevenshteinNorm}}("dairy", "diary") = \frac{2}{\max(5, 5)} = \frac{2}{5} = 0{,}4$$

Evaluation (7) : Instanciation/analyse des res.

Mesures de Similarité et de Distance

- **Autres distances :**
  - **Distance de Jaro**
    - *Basée sur le nombre de caractères correspondants entre deux chaines.*
    - *Deux caractères peuvent être considérés comme correspondants si leurs positions ne sont pas trop éloignées (seuil d'éloignement).*
  - **Distance de Jaro-Winkler**
    - *Idem + les chaines ayant des préfixes similaires sont favorisées.*