SPELL-CHECKING QUERIES BY COMBINING LEVENSHTEIN AND STOILOS DISTANCES

Zied Moalla[1, 2], Lina F. Soualmia[1, 3], Élise Prieur-Gaston[1]
Thierry Lecroq[1], Stéfan J. Darmoni[1]

[1] CISMeF, Rouen University Hospital & TIBS, LITIS EA 4108, University of Rouen, France
[2] MIRACL, Sfax University, Tunisie
[3] LIM&Bio EA 3969, Sorbonne Paris Cité, France

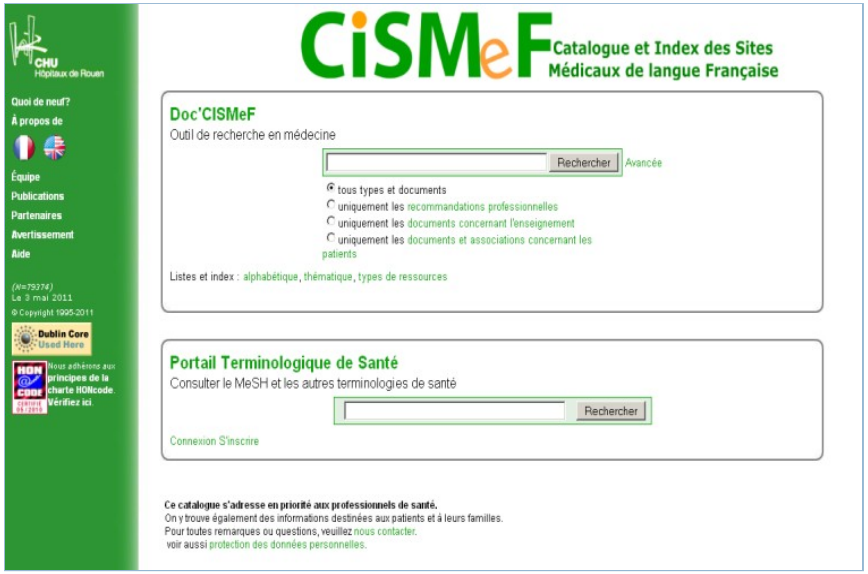
NETTAB 2011

Clinical Bioinformatics
October 12-14, 2011, Pavia, Italy

Content

- Context
- Introduction
- Materials and methods
  - Levenshtein distance
  - Stoilos distance
- Results
- Conclusion
- Perspectives

Context



Catalog & Index of Health Resources in French on the Internet

CISMeF = quality controlled health gateway for French institutional health resources

Doc'CISMeF: a search tool

- to search within the catalog CISMeF more than 82,000 documents
- specific of the health resources available on the Internet, such as association, patient information, community networks

3 types of users:
- Patients
- Students
- Clinicians

Introduction

- Increase in the number of users querying different search engines
- Internet became a major source of health information
- Medical vocabularies are difficult to handle by non-professionals

- "*Did you mean:*" *of Google or* "*Also try:*" *of Yahoo*

Introduction

- **Purpose:** Spelling correction for medical queries in French.
- **Method:** Spelling correction based on comparing the query with a dictionary.
- **Tools:** The string distance of Stoilos and the Levenshtein edit distance to correct spelling errors. We propose here to combine them.

String distance: Levenshtein

- Minimum number of edit operations (insertion, deletion, substitution) to transform one string into the other

String distance: Levenshtein

- The Normalized Levenshtein (*LevNorm*) in the range [0, 1]

$$LevNorm\ (c_1, c_2) = \frac{Lev\ (c_1, c_2)}{\text{Max}\ (length(c_1), length\ (c_2))}$$

- Example :

LevNorm (Trigonocepahlie, Trigonocephalie) = 2/15 = 0.133

Lev(Trigonocepahlie, Trigonocephalie) = 2

max(*length*(Trigonocepahlie), *length*(Trigonocephalie)) = *max*(15,15) = 15

String distance: Stoilos

- The similarity among two entities is related to their commonalities as well as to their differences. Thus, the similarity should be a function of both these features.

$$Sto(s_1, s_2) = Comm(s_1, s_2) - Diff(s_1, s_2) + Winkler(s_1, s_2)$$

String distance: Stoilos

- The function of commonality computes the longest common substrings between 2 strings

$$Comm\left(s_1, s_2\right) = \frac{2 \times \sum_i length\left(MaxComSubString_i\right)}{length\left(s_1\right) + length\left(s_2\right)}$$

- Example: s_1= 'Trigonocepahlie' et s_2= 'Trigonocephalie'

  $length(MaxComSubString_1)$ = $length$(Trigonocep) = 10

  $length(MaxComSubString_2)$ = $length$(lie) = 3

  $Comm$(Trigonocepahlie,Trigonocephalie) = 13/15 = 0.866

String distance: Stoilos

- Based on the length of the unmatched strings that have resulted from the initial matching step

$$Diff(s_1, s_2) = \frac{uLen_{s_1} \times uLen_{s_2}}{p + (1-p) \times (uLen_{s_1} + uLen_{s_2} - uLen_{s_1} \times uLen_{s_2})}$$

s_1 = 'Trigonocepahlie' and s_2 = 'Trigonocephalie' and p = 0.6

$uLen_{S1}$ = 2/15 and $uLen_{S2}$ = 2/15

So $Diff(s_1, s_2)$ = 10/787 = 0.0254

String distance: Stoilos

- The Winkler correction:

$$Winkler\left(s_1,s_2\right) = L\times p'\times\left(1-Comm\left(s_1,s_2\right)\right)$$

s_1 = 'Trigonocepahlie' and s_2 = 'Trigonocephalie'

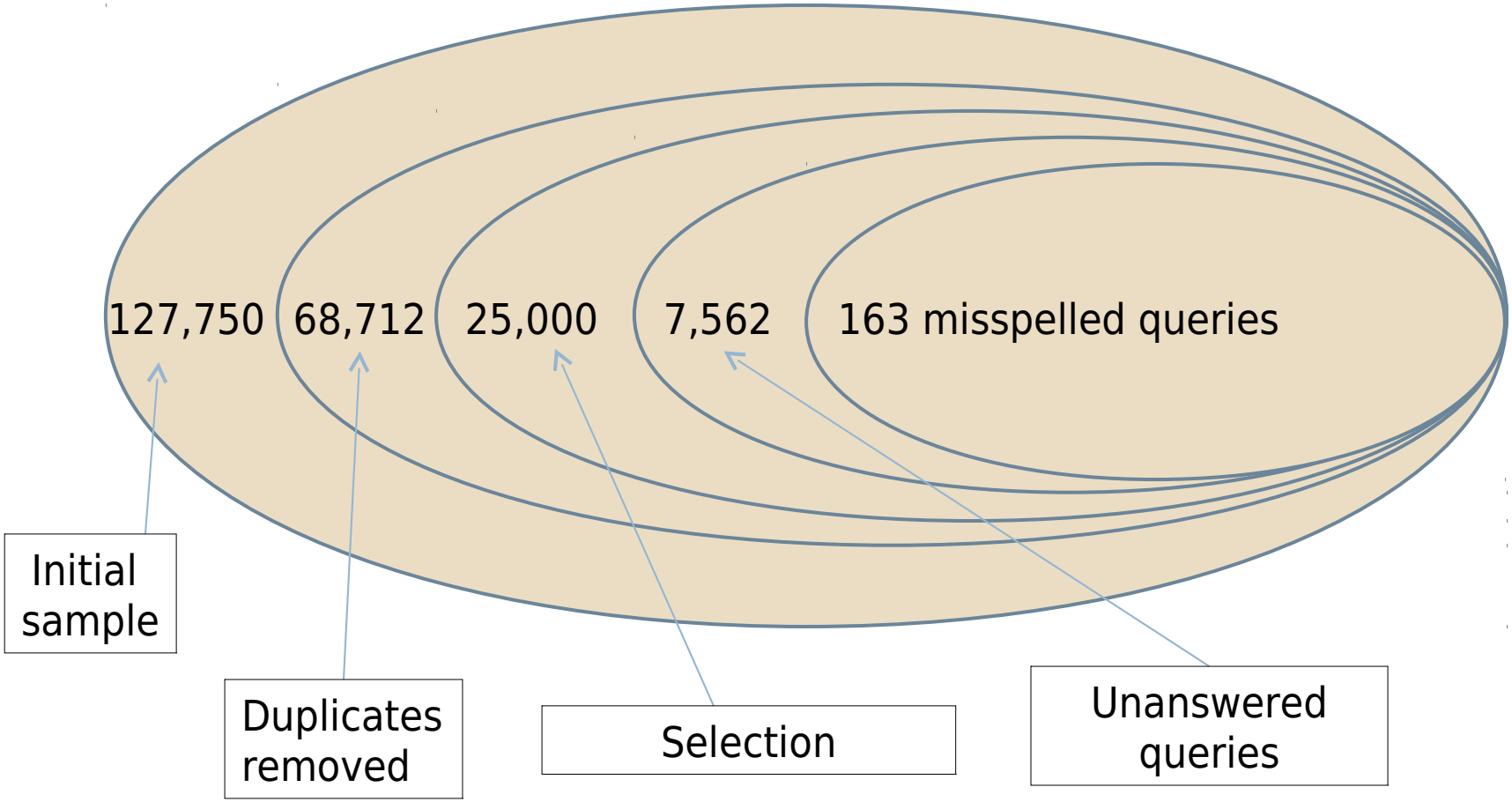
$L = 4$ and $p' = 0.1$

So $Winkler(s_1,s_2) = 4/75 = 0.053$

- Altogether

Sto(Trigonocepahlie, Trigonocephalie) = 13/15 – 10/787 + 4/75
= 0.894

Materials: Queries



Choice of thresholds

Levenshtein and Stoilos string distances require a choice of thresholds to obtain a manageable number of propositions of correction to the user.
So we have tested this number for 163 misspelled queries.

	Method							
	Levenshtein			**Stoilos**			**Levenshtein & Stoilos**	
Thresholds	<0.2	<0.1	<0.05	>0.7	>0.8	>0.9	Lev < 0.2 Stoilos > 0.8	Lev < 0.2 Stoilos > 0.7
Nb of answers	224 1.37	76 0.46	8 0.04	1454 8.92	489 3	140 0.85	179 1.09	213 1.30

Evaluation

$$\text{Recall} = \frac{\text{Queries correctly corrected}}{\text{Queries}}$$

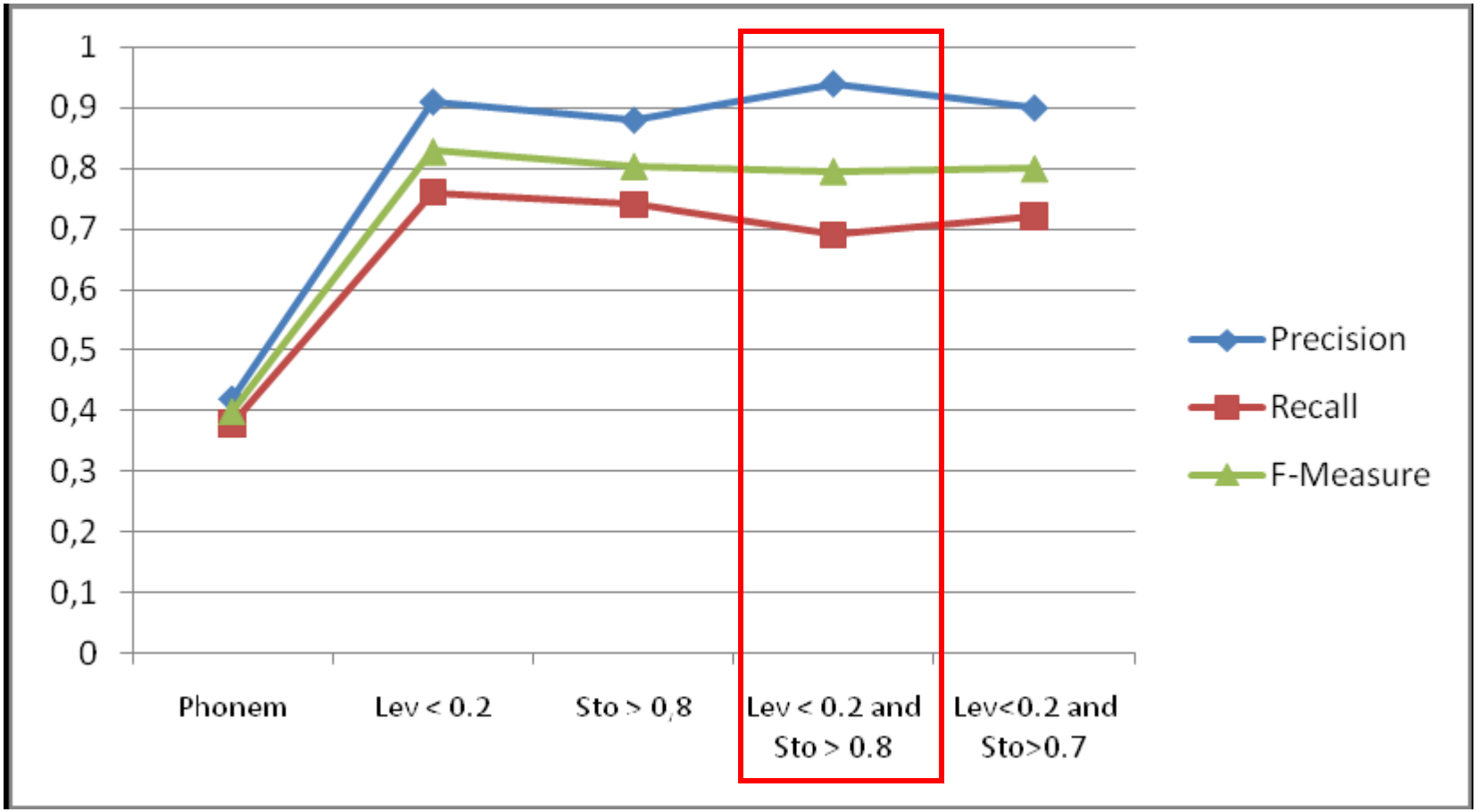
$$\text{Precision} = \frac{\text{Queries correctly corrected}}{\text{Queries corrected}}$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Results

Method	Recall	Precision	F-Measure
Phonetic transcription	0.38	0.42	0.399
Levenshtein < 0.2	0.76	0.91	0.8283
Stoilos > 0.8	0.74	0.88	0.8039
Levenshtein < 0.2 & Stoilos > 0.8	0.69	0.94	0.7958

Evaluation


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Phonem
Lev < 0.2
Sto > 0,8
Lev < 0.2 and Sto > 0.8
Lev<0.2 and Sto>0.7
Precision
Recall
F-Measure

Conclusion

- A method to automatically correct misspelled queries submitted to health search tool
- The combination of the 2 distances gives a recall of 69% and a precision of 94%
- This combination has increased the precision, but decreased the recall
- The functionality is implemented in CISMeF

Perspectives

- Misspelled queries categorized according to their number of words
- The configuration of a keyboard, by studying the distances between keys