Hosur'Tech participation to interactive INFILE

John Anton Chrisostom Ronald, Aurélie Rossi, Christian Fluhr
Cadege/Hossur'Tech
Contact: {Ronald.chrisostom, rossi.aurelie, christian.fluhr}@gmail.com

Abstract:

Tasks performed: interactive InFile French to French and English

Main objectives of experiments: As Hossur'Tech started from scratch in mid January to build an information extraction system based on a deep linguistic analysis, InFile runs were too early to be able to use our linguistic tools. Our objective in performing runs was to experiment comparison methods on real data to help us to design our future system.

Approach used: topics have been processed using a limited version of XFST with our own resources. Part of speech tagging and lemmatization were obtained. For the documents, it was not possible to use the same linguistic processing because of volume limitation of our version of XFST. A simple dictionary look-up without disambiguation was used. We were only able to process French and English in time. Arabic needed a little more time.

For each topic their title, description, and narrative contents were used. The example document was only used as a first positive feedback but not included strictly in the topic. For documents only title and text were used.

All document words inferred monolingual equivalents (for French to French comparison) or translations (for French to English comparison).

A word intersection was computed and then a concept intersection was established. All words inferred from the same word were considered as representing the same concept.

Each concept contained in the topic-document intersection receives a weight according to both a statistics computed on a similar corpus (Clef corpus) and the fact that the concepts are in the topic keyword list or title or not.

Proper nouns receive also an increased weight.

A tentative threshold between relevant and irrelevant documents was computed between the weight of the example document and the maximum weight of documents relevant to other topics.

Adaptation: The threshold has been adjusted according to the simulated feedback. Each word included into >= 2 relevant documents are included into the topic word set. We have asked 4 feedbacks for each topic which is too small according to real use of such systems.

Resources employed: own dictionaries

Results obtained: a great number of non relevant documents due to the fact that the feedback did not permit to adjust the threshold. The fact that we have not considered that documents could have several topics has also produced a large number of irrelevant documents. The low level of feedback for each topic (4) was not enough to add words from relevant documents in topics.

ACM categories and subject descriptors: H.3.3 Information Search and Retrieval, Information filtering

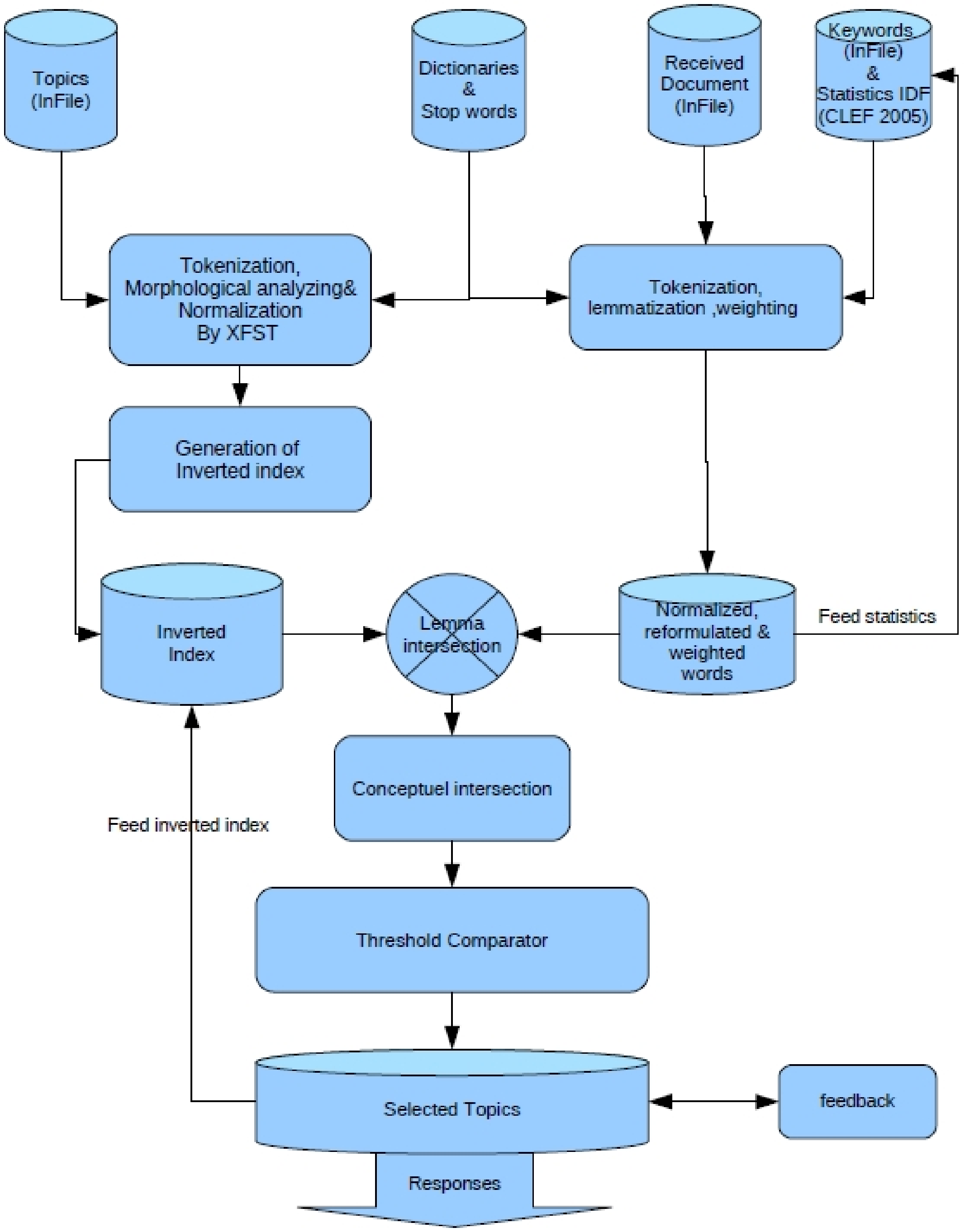
Free keywords: adaptive filtering, cross-lingual filtering, natural language processing

1 Introduction:

Hossur'Tech started from scratch in mid January 2009 to build an information extraction system base on a deep linguistic analysis. We have decided to base our linguistic processing upon weighted finite state automatons. For this purpose we are developing a language to build multilingual linguistic processing based on the openFST framework.

The planning of this technology development was not compatible with the participation to INFILE. We were conscious that INFILE participation and INFILE test data afterwards are a very valuable means for designing and tuning our future system. We have decided to participate but with much less elaborated linguistic processing. The main objective was to study the comparison strategies, weighting of intersections and finding a threshold to discriminate relevant and irrelevant documents in a context where statistical discriminating tools cannot be applied.

2 Functional diagram :



This functional diagram sum up the whole process of our system.

3 Linguistic processing:

A same linguistic processing must be used both on topics and on documents. In our case it was not possible. Waiting for our new technology, we decided to use the Xerox XFST automaton compiler to develop a morpho-syntactic parsing based on our existing language resources. The available version of XFST has limitations that are not annoying to process limited amount of text like topic texts but prevent to process a large corpus of documents like the one used in Infile.

For topics we processed all the fields: title, descriptive, narrative and keywords using the full parser (part of speech tagging and compound recognition).

For the documents we processed headline and data content. As XFST cannot be used we have developed a simple dictionary look-up giving all the possible lemmatizations without disambiguation. As the dictionary used by XFST and the look-up for documents are the same, intersection can be obtained.

Before processing, accents and all punctuations including hyphens were removed. For the topics which were treated with accents, they were removed during the comparison processing.

4 The problem of interactive adaptive filtering:

To categorize a document between two values (relevant and irrelevant) there is a lot of methods but generally based on a learning of positive and negative examples.

These methods cannot be used in our case because at the beginning we have only one example of a positive document. 200 feedbacks for 50 topics are not enough (4 for each topic) to have a good learning sample. In fact 4 feedbacks for one topic is not real life case. A user can give more feedbacks and also we can infer simulated feedbacks by observing his reading of proposed documents.

We have two kind of problems to solve: first to elaborate a concept intersection between a topic and an arriving document and to weight this intersection to get a relevance weight.

The second problem is to choose a threshold under what, documents are considered as irrelevant. Choosing a high threshold can lose relevant documents. Choosing a low threshold can give a lot of irrelevant documents that cannot be corrected by only 4 feedbacks for each topic.

5 Choices for our run:

Computation of the word intersection:

All document words infer equivalents in the same language using synonyms and other monolingual thesaurus-like reformulation rules for French to French comparison.

Example: lutte → combat, bataille, dispute

All document words infer equivalents in the other language using bilingual reformulation rules for French to English comparison.

Example: fight → combat, lute, dispute

All lemmatized topic words are organized into an inverted file. A special procedure computes the best intersection between the arriving document and all the topics.

Weighting of the words:

All words do not bring the same information (discriminative power). As it is not possible to have a statistics on a corpus which is not yet received, we have computed a general weight (based on idf) on a similar corpus (clef 2005).

We have also considered the importance of Keywords and gave them a better weight than other topic words.

We have also increased the weight of proper names. This was a conclusion of CEA in TREC 8 that increasing of proper names weight increase the performance of the system. In fact, filtering systems are often used to track persons, or companies or places.

Weighting of the intersections

Topic words and inferred words from topic words do not constitute a big set of words. Links between original topic words and inferred ones are kept. We will now consider not the word intersection but the concept intersection. The original topic words and all the words inferred from it are considered as equivalent to represent a concept.

The weight attributed to a concept is the minimum weight of all the words representing the concept.

Weights are added to give a relevance value to the intersection (ie to the topic)

6 Computation of a first threshold:

As we have a relevant document in the topic, we compute the value of its intersection with the topic. This gives an upper value for the threshold.

To obtain examples of irrelevant documents we have considered that all the example documents attached to other topics are irrelevant documents. In somes cases when topics have some intersection between each other, this can be a good way to discriminate topics.

We have chosen the greater value for the irrelevant examples as a lower limit for the threshold.

The threshold used at the beginning was set at lower value +85% of the difference between lower and upper threshold

7 Adaptation:

Three kind of adaptation have been used.

The first concerns the threshold. If a positive feedback is given, and if the conceptual intersection value is lower than the previous upper threshold, the new one is used. If a negative feedback is given and the value is upper than the previous lower threshold, the new one is used.

The second adaptation is devoted to add relevant vocabulary to the topic. As the fact that a word is in one relevant document is not a strong reason to consider it as a word relevant for the topic, we have considered, to eliminate hazard, that the presence of a word in two documents attested as relevant is necessary to add it into the topic word set.

The weight of words (based on idf) are updated,, new words are added by the processing of entering document.

8 Results and discussion:

Ours results show that we have produced a too large set of irrelevant documents.

A first study of our results shows that the fact that compounds are not used has facilitated the production of irrelevant documents. This is especially strong in these particular topics where elements of compounds are widespread words. For example, we have used “économie (economy)” and “sport” whose use is widespread instead of “économie du sport (economy of sport)” which is more focused. This will be eliminated with our full linguistic analysis.

Another cause of this bad precision is due to the fact that we have considered intersection at the document level. A lot of documents are multi topic documents. With this kind of documents and without the use of compounds, general purpose words can intersect the documents on various parts without any significance.

To solve this problem, we are introducing a computation of intersection based on passages. Only passages that concentrate topic words (or inferred words from topics words) are considered for computing the intersection.

We have lost relevant documents.

Topics are a little bit general. It is necessary to infer more specific words like those that can be found in the not found documents. To solve this problem using an automatic way, it is necessary to have a larger feedback that can be used to learn associated words. And even in this case, the user will lose a lot of relevant document at the beginning.

In real applications either the domain in which topics are designed is important enough to build an ontology that can be used to expend the topics or the user is asked to elaborate a larger description as topic giving all the vocabulary needed to find a maximum of relevant documents.

Example of domain ontology: IAEA has build for the TOPIC system of Verity (now Autonomy) an ontology to follow nuclear proliferation activities.

Example of what the user can do for the topic on “compétition sportives internationales”. He can add: coupe du monde, coupe d’Europe, coupe d’Afrique, championnat du monde, jeux olympiques ,… Without that it is difficult to find relevant documents especially at the beginning of the filtering.

9 Conclusion:

If the InFile interactive track is a very difficult task, to have real data and real conditions of use is very important to develop operational systems. We will use the evaluation package to develop our system. Even if our results are now quite bad this experience will be very useful to produce a good operational system when all basic tools and especially the linguistic processing ones will be available.

10 Bibliography:

- EMIR at the CLIR track of TREC7, F. Bisson, J. Charron, C. Fluhr, D. Schmit, November 9-11 1998, Gaithersburg, MD
- The TREC 2001 Filtering Track Report, Stephen Robertson (Microsoft Research Cambridge ,Uk), Ian Soboroff (NIST, USA), Nov. 2001, Gaitherburg, MD

Guidelines for Full Papers

The text of the full papers should respect the following guidelines.

size: A4

format: pdf (if difficult, ps or MS Word (rtf) are acceptable) Do NOT lock the pdf file.

borders: top, left right 2.5 cm ; bottom 3 cm

text size: 16 x 24 cm

length: We suggest 10 page maximum but as we are publishing electronically this is not too important

title: Times 14 pt bold centered

author(s): Times 10 pt centered

abstract: Times 10 pt justified (abstracts should provide main details of paper, see above)

ACM Categories and Subject Descriptors: Times 10 pt left aligned

Free Keywords: Times 10 pt left aligned

body text: Times 10 pt justified

Section Headings: Times 12 pt bold left aligned

Emphasis: Times 10 pt italic