STAM : traduction des textes non structurés (dialectes du Maghreb)

Mehdi Embarek[1, 2] Soumya Embarek[1]
(1) MK SOFT, 20 Rue des Réservoirs, 94800 Villejuif, France
(2) MED POINT DZ, 30 Rue du Hoggar, 16035 Alger, Algérie
embarekm@gmail.com, s.embarek@mksoft-fr.com

RÉSUMÉ

L'utilisation des plateformes de communication (réseaux sociaux, forums de discussions, ...) a pris une ampleur considérable. Ces plateformes permettent aux internautes d'exprimer leur avis concernant un sujet, demander ou échanger des informations, commenter un événement, etc. Ainsi, nous retrouvons dans ces différentes sources d'informations une quantité importante de textes rédigés dans des dialectes locaux dont sont originaires les rédacteurs. Cependant, ces textes non structurés rendent l'exploitation des outils de traitement automatique des langues très difficile. Le système STAM aborde cette problématique en proposant un système capable de transcrire automatiquement des textes écrits dans un dialecte parlé dans les pays du Maghreb en un texte facilement interprétable et compréhensible (français ou anglais).

ABSTRACT

STAM: Translation of unstructured text (Maghreb dialects)

The use of communication platforms (social networks, discussion forums...) has grown considerably. These platforms allow users to express their opinions about a subject, ask or exchange information, comment on events, etc. Thus, we find in these different sources of information a large amount of texts written in local dialects of origin of the editors. However, these unstructured texts make the use of the language processing tools very difficult. The STAM system addresses this problem by providing a system able to automatically transcribe texts written in a dialect spoken in the Maghreb countries in an easily interpretable and understandable text (French or English).

MOTS-CLÉS : Dialecte, Arabe, Maghreb, Multi-dialectes, Transcription, STAM.
KEYWORDS: Dialect, Arab, Maghreb, Multi-dialects, Transcription, STAM.

L'évolution de l'usage d'Internet et le développement de nouveaux langages et méthodes de programmation a complètement changé le mode de communication des internautes. De nos jours, ces derniers utilisent différentes plateformes (réseaux sociaux, forums de discussions, tweets, etc.) pour exprimer leur avis concernant différents sujets, demander ou échanger des informations, commenter un évènement, etc. Aussi, nous retrouvons dans ces différentes sources d'informations des textes rédigés dans des langages informels. Ces langages représentent généralement les dialectes locaux dont sont originaires les rédacteurs, ce qui rend la compréhension du texte pratiquement impossible pour les personnes ne parlant pas les dialectes employés. Cette quantité importante de textes non structurés peuvent être considérés comme des sources d'informations et il serait intéressant de pouvoir les exploiter et les analyser afin d'en extraire le contenu informationnel en se basant sur des outils et techniques adaptés.

Dans le monde arabe, et plus particulièrement dans les pays du Maghreb, on recense plusieurs dialectes qui diffèrent d'un pays à l'autre. Aussi, dans un même pays, plusieurs dialectes peuvent exister selon les régions du pays. Ces différents dialectes sont une association de l'arabe littéraire et du berbère (langue autochtone du Nord de l'Afrique) influencés par d'autres langues étrangères telles que le français, l'italien ou l'espagnol. En effet, les mots de ces langues étrangères ont intégré ces dialectes ; certains ont été phonologiquement modifiés pour s'adapter à la structure de la langue arabe mais d'autres sont utilisés en tant que tels sans aucune modification. De ce fait, pour un texte en arabe dialectal, il n'existe aucune règle d'écriture, un mot peut éventuellement être écrit de plusieurs manières. Par exemple, le mot «restaurant» (resto, erresto, mataam, mat3am, mete3em, …). On note ici la présence de chiffres dans certains mots pour exprimer principalement une certaine prononciation (problème phonétique) spécifique aux mots arabes. Cette particularité rend l'utilisation des outils de traitement automatique des langues très difficile, notamment pour la traduction des textes exprimés en arabe dialectal. Un autre point crucial concernant le traitement de l'arabe dialectal concerne le manque de ressources (terminologies) exploitables pour les dialectes. De nombreux travaux ont été menés dans ce sens afin d'étudier plus particulièrement les spécificités des dialectes (Guella, 2011) (Meftouh et al., 2012) sans pour autant aller plus loin dans le développement d'outils adaptés à l'exploitation des textes.

Le projet STAM (Système de Transcription AutoMatique) (http://www.stam-dz.com) aborde cette problématique en proposant un système capable de transcrire automatiquement des textes écrits dans un dialecte parlé dans le monde arabe (langue source) en un texte facilement interprétable, compréhensible et en bon français ou anglais (langues cibles). Cette transcription est également valable dans l'autre sens. STAM offre ainsi la possibilité de pouvoir interpréter et extraire les informations contenues dans les textes traduits en utilisant les outils du TAL (recherche d'information, traduction, filtrage d'information, résumé automatique, etc.).

Dans sa première version (Embarek, 2014), le système STAM ne permettait que de transcrire les textes issus du dialecte algérien. Grâce à l'enrichissement de sa terminologie multi-dialectale et de ses règles d'écriture, la nouvelle version de STAM prend en compte de nouveaux dialectes : le dialecte amazigh (berbère), le dialecte égyptien, le dialecte marocain et le dialecte tunisien. Aussi, le système permet d'interpréter les textes (en dialecte) directement en anglais contrairement à la précédente version qui ne proposait que le français comme langue cible.

Enfin, la prochaine étape dans le développement du système STAM consiste à enrichir sa terminologie. Cet enrichissement va permettre d'améliorer la couverture des différents dialectes étudiés dans les textes.

Références

EMBAREK M. (2014). Le système STAM. *TALN 2014*, 21-22.

GUELLA N. (2011). Emprunts lexicaux dans des dialectes arabes algériens. *Synergies Monde Arabe 8*, 81-88.

MEFTOUH K., BOUCHEMAL N., SMAÏLI K. (2012). A study of a non-resourced language : an algerian dialect. *The Third International Workshop on Spoken Languages Technologies for Under-resourced Language*, 125-132.