

La traduction automatique pour l'entreprise

Livre blanc

Qu'est-ce que la traduction automatique ?

Le terme de traduction automatique désigne une technologie qui permet de traduire automatiquement par ordinateur un texte d'une langue vers une autre, dans le but de le rendre compréhensible aux locuteurs de la langue de destination.

Pourquoi la traduction automatique ?

Les raisons pour une entreprise de recourir à la traduction automatique sont multiples :

- apparition de nouveaux contenus dans l'environnement professionnel,
 - entreprise en voie de globalisation avec de nouveaux interlocuteurs internationaux,
 - nombreux contenus dont on a besoin de saisir le sens sans nécessairement en saisir tous les détails (« gisting »),
 - intranet régulièrement mis à jour,
 - contenu hautement confidentiel ne pouvant être externalisé pour traduction,
- parmi d'autres.

On peut en général classer les usages de la traduction automatique en deux grandes catégories : assimilation (comprendre un texte dans une langue que l'on ne connaît pas) et dissémination (communiquer avec des locuteurs de langues différentes).



La traduction automatique basée sur des règles (rule-based, RBMT)

Technologie de traduction automatique apparue historiquement la première, utilisant des règles grammaticales d'analyse et de transformation, associées à des dictionnaires bilingues ou multilingues, afin d'élaborer un moteur de traduction. Les applications qui l'utilisent comprennent en général des outils comme des éditeurs de dictionnaires et de règles, permettant de personnaliser les traductions. D'une conception basée sur la grammaire, elle est relativement simple à mettre en œuvre, et son optimisation passe avant tout par la construction de dictionnaires personnalisés ciblant le ou les domaines métier de l'entreprise. L'optimisation des outils RBMT se fait également par l'élaboration de règles syntaxiques ou de normalisation du texte. De ce fait, si la traduction automatique basée sur les règles produit en général des traductions plutôt littérales, la terminologie et la structure grammaticale des phrases employées y sont scrupuleusement respectées. Parmi les concepteurs de ce type de technologie on trouve notamment Systran, ProMT, Linguatrec.

Les grands types de traduction automatique

La traduction automatique statistique

Technologie de traduction automatique utilisant des algorithmes et d'importants corpus bilingues pour élaborer un moteur de traduction. Les règles de traduction sont générées par calcul (« pattern matching ») au terme d'un processus d'analyse des corpus bilingues appelé entraînement. La qualité du résultat est en relation directe avec la qualité des corpus sur lequel le moteur est entraîné : grammaire et orthographe correctes, adéquation du contenu du corpus vis-à-vis des textes à traduire, style, cohérence terminologique

L'optimisation de tels systèmes passe par l'ajout et l'amélioration de nouveaux corpus spécialisés (« in-domain »), ce qui peut paraître de prime abord plus simple qu'un système basé sur les règles où il s'agit d'ajouter des mots au dictionnaire, mais il demeure plus complexe à paramétrer dans le détail, et ces corpus doivent aussi être préparés spécifiquement pour cet usage. Son mode de fonctionnement amène logiquement une traduction stylistiquement plus fluide, mais grammaticalement plus hasardeuse et avec une terminologie qui n'est pas toujours prévisible.

Il existe des moteurs déjà entraînés sur des corpus génériques ou multi-domaines, accessibles en ligne lorsqu'ils sont ouverts au public, selon différentes modalités commerciales ou de sécurité, et des systèmes à construire soi-même à partir d'éléments open source comme Moses pour un usage propriétaire.

La traduction automatique statistique est représentée commercialement notamment par Google Translate, Morphologic ou IBM WebSphere.

Les traductions automatiques hybrides : convergence des technologies

Depuis quelques années sont apparues des technologies dites « hybrides », qui tentent de prendre le meilleur de chacune des deux technologies décrites précédemment. Elles peuvent, selon les cas, être conçues sur une technologie basée sur des règles à laquelle est ajoutée une couche de correction automatique de type statistique, utilisant de larges corpus, ou bien consister d’abord en une technologie statistique dans laquelle sont instillés des éléments d’analyse grammaticale. Systran, Microsoft Translator (Bing) et AppTek font partie des tenants de cette perspective.

	TA basée sur des règles	TA statistique
Principes de base	règles d’analyse syntaxique règles de transformation dictionnaires multilingues	corpus pattern matching
Mise à jour	ajout de termes aux dictionnaires ajout de règles	ajout de nouveaux corpus préparés et validés entraînement
Traduction	littéralité cohérence de la terminologie présente dans les dictionnaires	fluidité cohérence et syntaxe variable

Produits open source

Il existe des produits « open source », permettant de construire son système de traduction propriétaire, pour les deux grands types de technologie.

On citera par exemple, Open Logos et Apertium pour la traduction automatique basée sur des règles. De son côté, la technologie statistique a beaucoup bénéficié du « kit » de développement Moses sur lequel beaucoup de produits professionnels se sont développés, comme ceux d'Asia Online, Kantan MT ou Safaba. Toutefois il s'agit dans sa version open source d'outils très généraux et de données brutes, qui demandent un développement conséquent, tant au niveau du moteur lui-même qu'au niveau de l'interface, pour aboutir à un résultat performant.

Produits de traduction sur Internet

La première chose qui vient souvent à l'esprit lorsque l'on parle de traduction automatique, sont les services disponibles gratuitement en ligne, et qui sont relativement bien développés comme Google Translate ou Microsoft Bing. Ces outils se présentent en général sous forme de fenêtres de saisies de texte, mais ils peuvent être disponibles également via des plugins dans la barre d'outils des navigateurs, ou des widgets directement sur les pages internet ou intranet. Non basés sur un domaine précis, mais sur l'ensemble des données disponibles sur l'internet, la qualité du résultat n'est pas prévisible. Cependant, des possibilités de personnalisation plus ou moins élaborées existent, selon les moteurs.

Produits pour LSP (Language Service Provider)

La plupart des produits de traduction automatique existants trouvent un débouché logique dans la fourniture de services de traduction automatique aux sociétés de traduction disposant des compétences techniques nécessaires pour les mettre en oeuvre. Ils leur permettent ainsi de construire des moteurs personnalisés destinés à leurs clients finaux, qui sont utilisés pour pré-traduire en traduction automatique des textes à post-éditer ensuite par des linguistes professionnels, permettant ainsi d'augmenter leur productivité, et de réduire les délais sur les volumes importants. Cela leur permet également de répondre aux besoins croissants de leurs clients finaux dans la diffusion des contenus à courte durée de vie à un coût de traduction compétitif.

Produits d'entreprise intégrés

Certains éditeurs ont développé leurs produits de manière à les intégrer dans les outils de l'entreprise (bureautique, courriel etc.), pour permettre aux utilisateurs travaillant à l'international de traduire par traduction automatique les textes qu'ils écrivent ou qu'ils veulent lire.

Toutefois, avec l'émergence du cloud computing, cette tendance va sans doute refluer au profit de l'intégration de la technologie linguistique dans des solutions SaaS.

Produits intégrés à la bureautique :

- Systran Server et Desktop Applications : plugins dans MS Office et dans les navigateurs Web.
- Microsoft Translator : plugin dans MS Office Word, widget insérable sur des pages Web.
- Reverso de Softissimo (basé sur ProMT), et sa barre d'outils pour navigateurs.

Autres solutions pour l'entreprise

Une autre solution pour l'entreprise consiste à mettre en œuvre, à partir de produits existants, une solution personnalisée composée en deux niveaux distincts :

- d'une part, le développement d'un ou plusieurs moteurs de traduction personnalisés parmi les technologies et les fournisseurs disponibles, selon les paires de langues requises, et le matériel existant chez le client (glossaires, mémoires de traduction, corpus de textes).

Ces moteurs personnalisés devant nécessairement s'accompagner de dispositifs de mesure du niveau de qualité de traduction obtenue et de sa pertinence vis-à-vis du besoin.

- d'autre part, l'utilisation de « connecteurs », pour intégrer les moteurs de traductions dans l'infrastructure de l'entreprise, là où se trouve le besoin : par exemple des plugins dans les applications de messagerie ou de bureautique, un widget sur les pages de l'Intranet, ou encore une fenêtre de saisie pour une traduction « à la volée » etc., selon les besoins spécifiques identifiés.

Construction de dictionnaires (dictionary building)

Processus d'élaboration d'un modèle de traduction basé sur des règles, utilisant des outils tels que l'extraction terminologique, l'extraction de mots inconnus ou la normalisation.

Corpus (corpus)

Ensemble textuel constitué d'un nombre important de phrase. Peut être monolingue ou bilingue.

Corpus d'entraînement (training corpus)

Corpus bilingue utilisé pour l'élaboration d'un moteur de traduction automatique statistique.

Corpus de test (testing corpus)

Partie du corpus d'entraînement utilisé pour calculer la qualité de traduction obtenue par comparaison entre la traduction de référence et la traduction automatique des mêmes phrases.

Corpus générique (out-of-domain corpus)

En traduction automatique statistique, corpus textuel bilingue général, destiné dans un entraînement, à servir de base pour la syntaxe et le vocabulaire courants.

Corpus spécialisé (in-domain corpus)

En traduction automatique statistique, corpus textuel bilingue relatif à un domaine précis, destiné dans un entraînement, à appliquer une terminologie et une phraséologie spécifique.

Dictionnaire de normalisation (normalization dictionary)

Dans un modèle de traduction basé sur des règles, glossaire monolingue utilisé soit pour corriger le texte source, soit pour modifier la traduction obtenue, selon la langue à laquelle il s'applique.

Dictionnaire de traduction (translation dictionary)

Dans un modèle de traduction basé sur des règles, glossaire bilingue ou multilingue comportant un encodage d'informations morphologiques et sémantiques.

Entraînement (training)

Processus d'élaboration d'un moteur de traduction automatique statistique à partir de corpus bilingues.

Modèle de langue (language model)

Élément logiciel constitué d'un ensemble de règles monolingues créée par un entraînement statistique et destiné à normaliser un texte d'une langue donnée.

Modèle de traduction (translation model)

Élément logiciel constitué d'un ensemble de règles syntaxiques ou statistiques bilingues destiné à traduire un texte d'une langue dans une autre.

Moteur de traduction (translation engine)

Voir Modèle de traduction

Normalisation (normalization)

Action de modifier un texte monolingue à partir de règles syntaxiques, généralement en vue de le simplifier ou de le rendre plus adapté à la traduction automatique.

Post-édition (Post-Editing Machine Translation – PEMT or PE)

Processus de correction et d'amélioration d'une traduction obtenue avec un système de traduction automatique.

Pré-édition (pre-editing)

Action de normalisation d'un texte source, effectuée automatiquement ou manuellement, en vue de le rendre plus adapté à la traduction automatique.

Traduction automatique (Machine Translation – MT)

Technologie permettant de traduire automatiquement un texte d'une langue donnée dans une autre langue.

Traduction automatique basée sur des règles (Rule-Based Machine Translation – RBMT)

Traduction automatique utilisant des règles de transformation et des dictionnaires multilingues pour construire un moteur de traduction.

Traduction automatique hybride (Hybrid Machine Translation – HMT)

Traduction automatique utilisant plusieurs technologies, tant basée sur des règles que statistique.

Traduction automatique statistique (Statistical Machine Translation – SMT)

Traduction automatique utilisant des algorithmes et d'importants corpus bilingues pour construire un moteur de traduction.

À propos de **Lexcelera**

Lexcelera, société de traduction fondée en 1986 à Paris, a toujours allié l'excellence de la traduction humaine à la productivité des technologies d'automatisation.

Pionnier dans la recherche et l'utilisation de la traduction automatique, elle est l'un des leaders des solutions d'ingénierie linguistique.

Lexcelera a également été la première société de traduction en France à recevoir la certification ISO 9001 :2000 pour la gestion de la qualité.