

Le concept clé de l'AIML : la réduction symbolique

Soumis par Philippe YONNET

05-01-2008

Dernière mise à jour : 05-01-2008

La {showdesc:réduction symbolique} est un concept un peu abstrait, mais qu'il est important de bien maîtriser pour parvenir à réaliser des fichiers AIML "élégants" : à la fois complets, mais comportant un nombre minimum de catégories. Définition

La réduction symbolique désigne, dans le contexte de l'AIML, toutes les techniques permettant de supprimer dans une forme (pattern) les éléments inutiles, pour ne garder que ce qui constituera un "stimulus" traitable par le chatterbot. Pourquoi opérer des "réductions symboliques"

Quelle que soit la langue (mais c'est surtout vrai en français), il existe de nombreuses manières grammaticalement juste de formuler une phrase ayant un sens donné. A ces variantes autorisées par la syntaxe, on peut aussi ajouter des variantes de graphies possible (Monsieur, M., Mr), et bien sûr, les variantes involontaires provoquées par les fautes de frappe, d'orthographe, de syntaxe, ou l'emploi d'une langue relâchée (ah les joies du langage SMS...)

La réduction symbolique a pour objectif de supprimer dans la chaîne de caractère initialement entrée par l'internaute, tout ce qui peut "masquer" un stimulus identifiable, et de "reduire" cette chaîne parfois très longue à un groupe de termes identifiable à l'aide d'une "forme" stockée dans une catégorie.

Les différentes étapes de la réduction symbolique

La réduction symbolique passe par plusieurs étapes, certaines gérées par l'analyseur syntaxique (parser) AIML, d'autres doivent être programmées par le botmaster en AIML.

La normalisation.

La normalisation consiste à "toiletter" la chaîne de caractères de tout ce qui peut empêcher une comparaison avec les "patterns" stockées dans les fichiers AIML. Il s'agit notamment d'éliminer les caractères bizarres, les espaces en trop, les caractères de ponctuation etc...

Ex :

Entrée : "Salut ma belle ! Dis-moi qui es ton botmaster ..."

Devient : "SALUT MA BELLE DIS-MOI QUI ES TON BOTMASTER"

L'élimination des mots vides

Certains mots n'apportent aucune information qui soit utilisable par l'analyseur syntaxique. Celui-ci ne cherche pas à déterminer le sens d'une phrase, mais des "formes", des suites de caractère... Dans ce contexte, ce qu'il faut repérer, ce sont des chaînes "caractéristiques", et non pas des termes génériques. On peut par exemple oublier sans problèmes les conjonctions et les adverbes et certains adjectifs

"DONC SALUT MA BELLE ET DIS-MOI VITE QUI ES TON BOTMASTER"

devient :

"SALUT MA BELLE DIS-MOI QUI ES TON BOTMASTER"

Le découpage en éléments simples

Si la phrase est longue, il est possible qu'elle contienne différents blocs qui peuvent (doivent) être analysés séparément. Certains parsers AIML autorisent le découpage à l'aide notamment des signes de ponctuation. Mais pas la plupart ne gèrent pas cette possibilité, en raison du manque de fiabilité de la ponctuation tapée par les internautes. Le découpage d'une phrase en éléments simples passe donc par des reconnaissances de forme successives, en cascade

Ex : "SALUT MA BELLE DIS-MOI QUI ES TON BOTMASTER" matche d'abord avec "SALUT *"

Après traitement de salut, il reste "MA BELLE DIS-MOI QUI ES TON BOTMASTER" à traiter.

Cette chaîne matche avec "* BOTMASTER"

Ce qui permet de générer la réponse appropriée suivante avec deux catégories AIML seulement :

Input : "Salut ma belle ! Dis-moi qui es ton botmaster ..."

Réponse : "Bonjour vous ! Mon botmaster s'appelle Philippe. Mais je ne vous en dirai pas plus, rester discrète fait partie de mon métier"

La standardisation

La standardisation consiste à gérer toutes les variantes d'un stimulus donné, afin de renvoyer toutes ces variantes vers une catégorie unique.

Par exemple :

HELLO = BONJOUR = HI = COUCOU = KIKOU => SALUT

Cela signifie que l'on peut créer une catégorie unique pour les salutations, capable de délivrer une centaine de réponses possible différentes. Cette réponse est activée, après réduction symbolique, par toutes les formes de salutations qui auront pu être entrées sous formes de catégories, et qui toutes renvoient vers la catégorie salutations.

La standardisation permet aussi de gérer les fautes de frappe, d'orthographe, d'accent etc...