Machine Learning

Natural Language Processing (NLP)

Traitement du langage : données

Classification :

- thème/genre (gutenberg.org : 57k livres)
- auteur (gutenberg.org : 57k livres)
- sentiment (Kaggle movie review : 222k commentaires rotten tomatoes)
- reconnaissance d'entités nommées (Kaggle Annotated Corpus for NER : 1.3M tags)
- ...

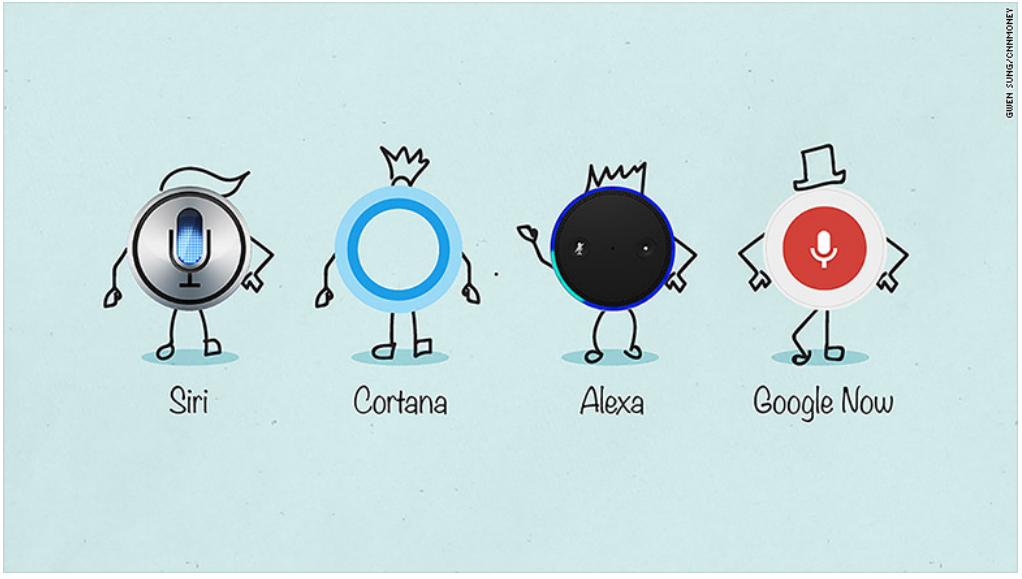
Traitement du langage : données

Compréhension :

- Question/réponses (SQUAD : 150k questions)
- Traduction (europarl : 450k phrases alignées)
- ...

Traitement du langage : Champs d'application

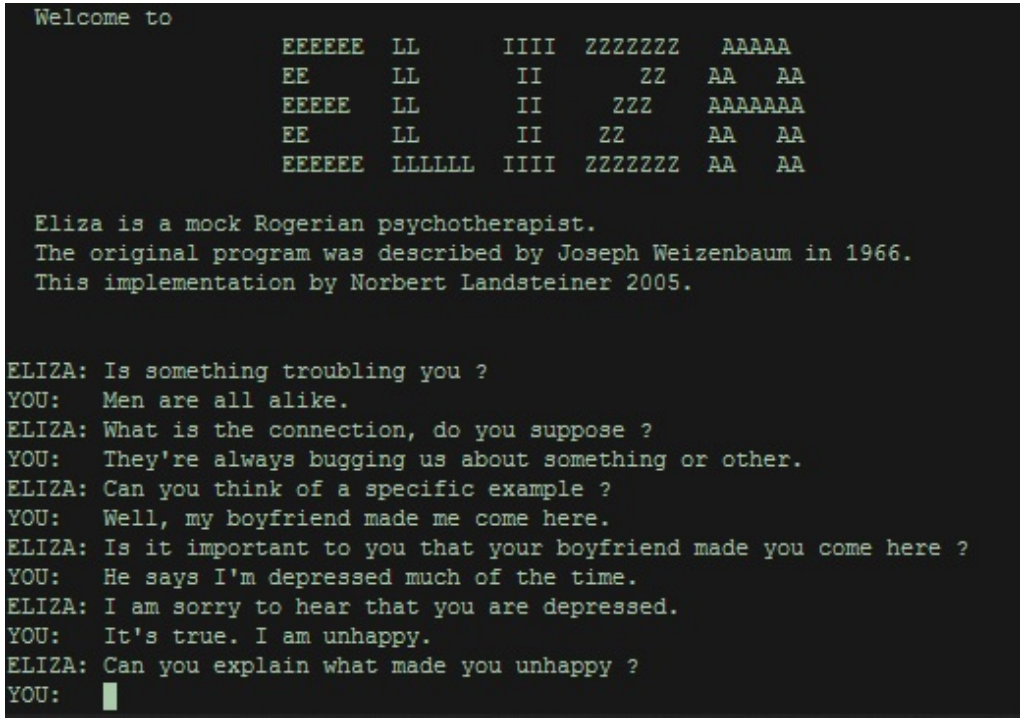
Transcription et synthèse de la parole



Identification du locuteur

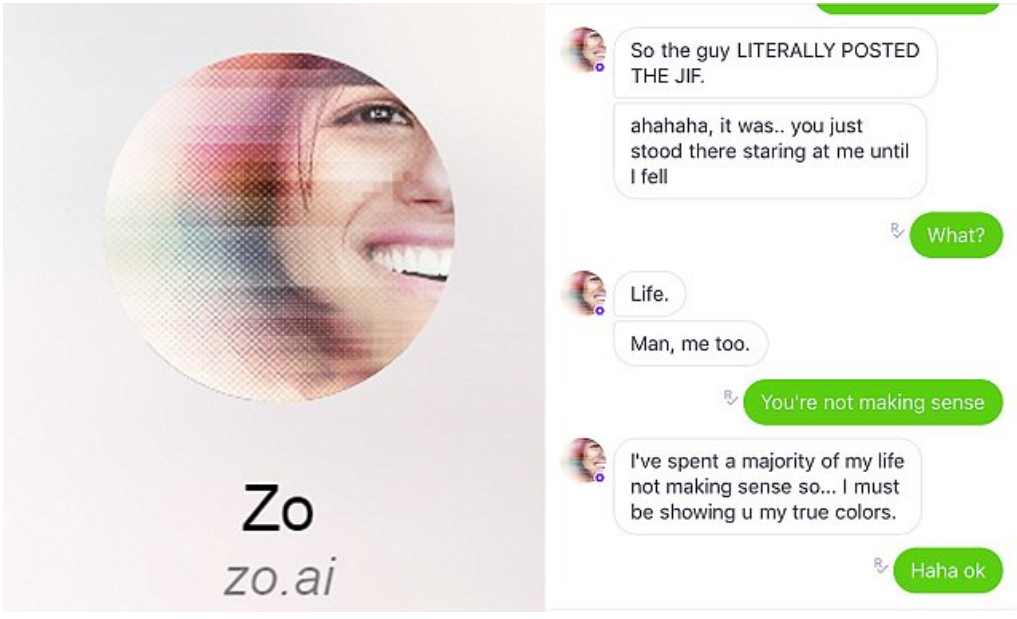
Traitement du langage : Champs d'application

Chatbot



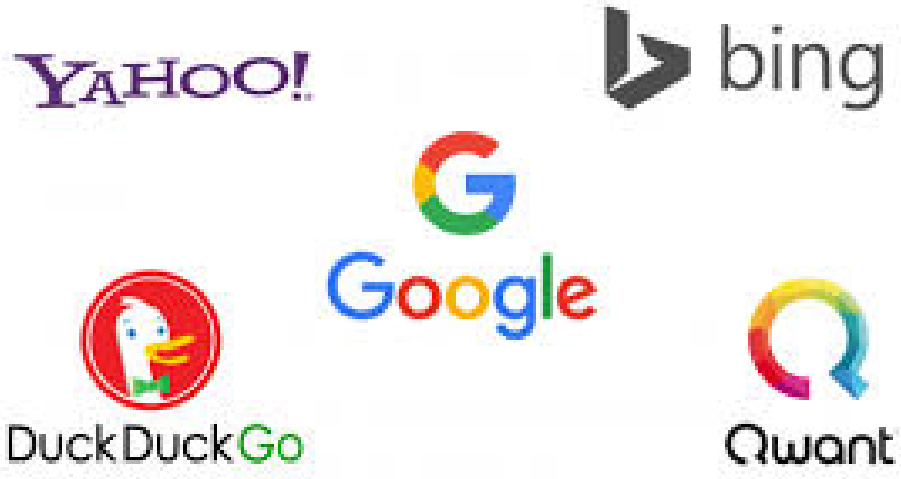
Traitement du langage : Champs d'application

Chatbot



Traitement du langage : Champs d'application

Moteur de recherche



Traitement du langage : Champs d'application

Extraction de données

Démo reconnaissance d'entitées nomées

Analyse de sentiments

Démo analyse de sentiments

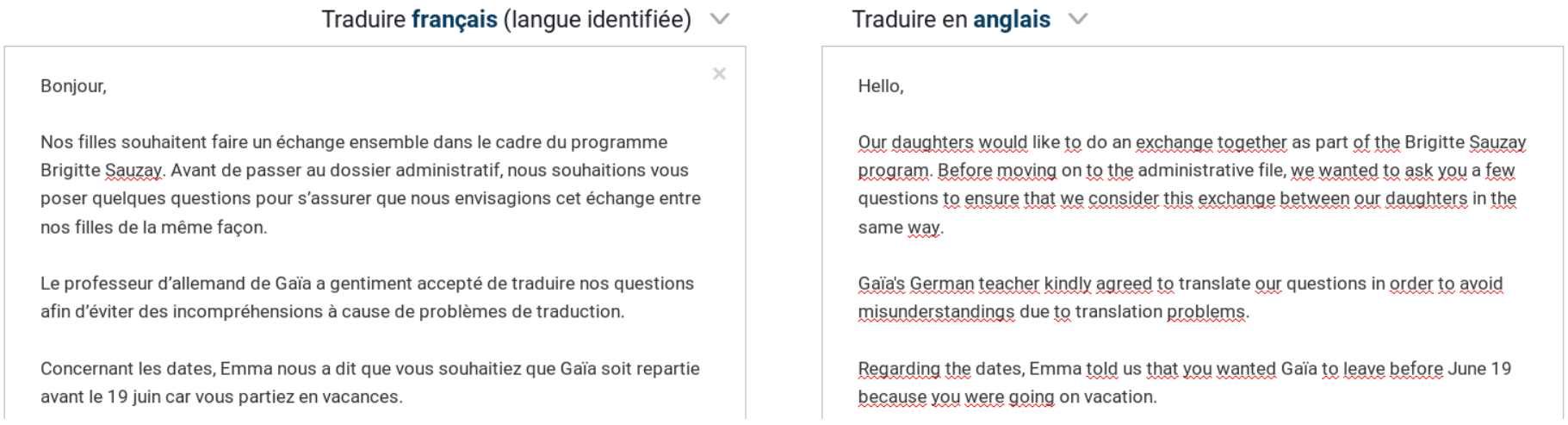
Traitement du langage : Champs d'application

Résumé

Input: Article 1st sentence	Model-written headline
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

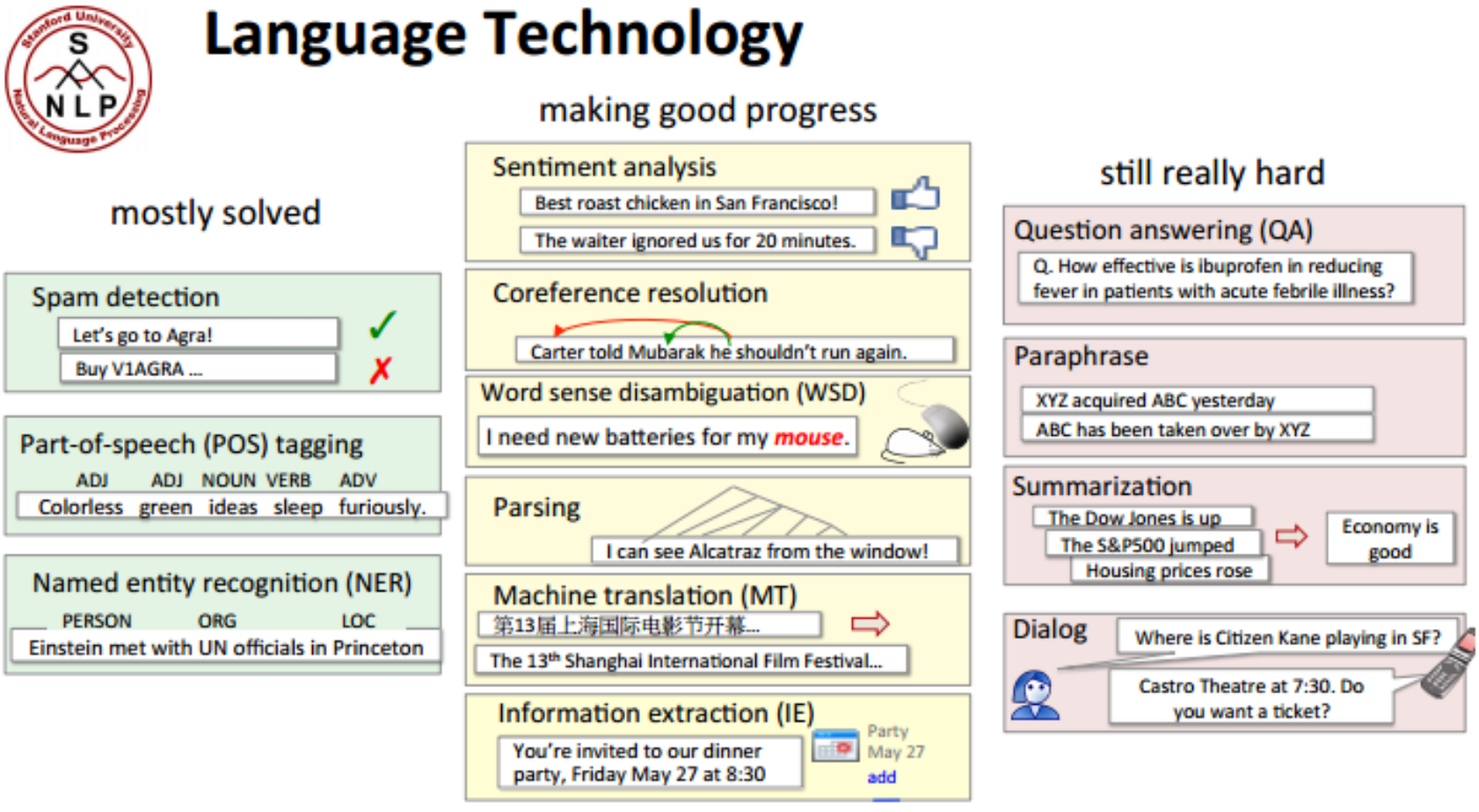
Traitement du langage : Champs d'application

Traduction



Démo traduction de DeepL

Traitement du langage : Champs d'application



Traitement du langage : WordNet

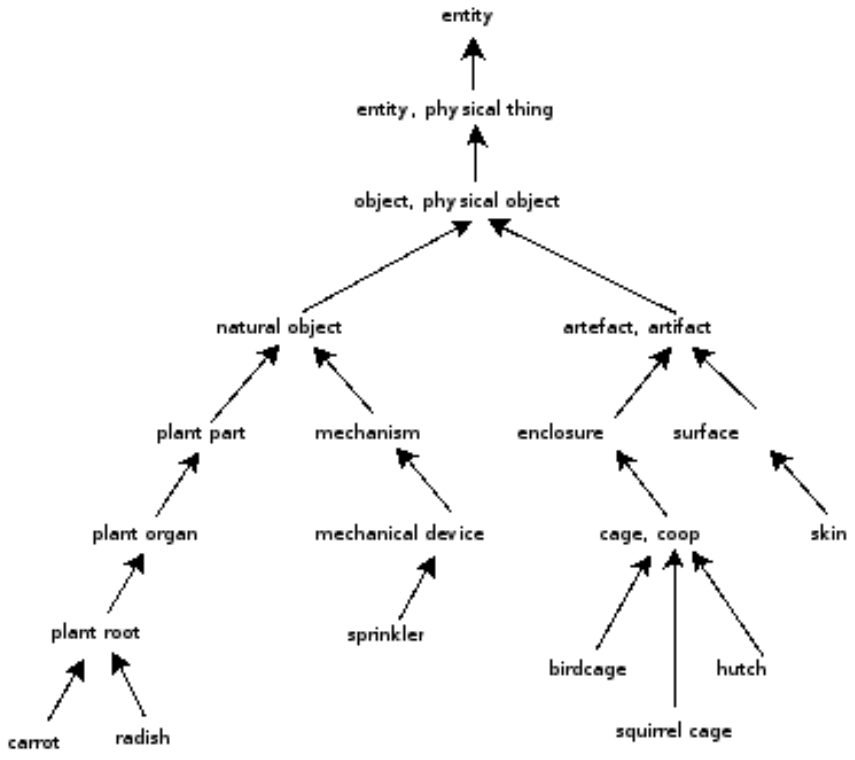
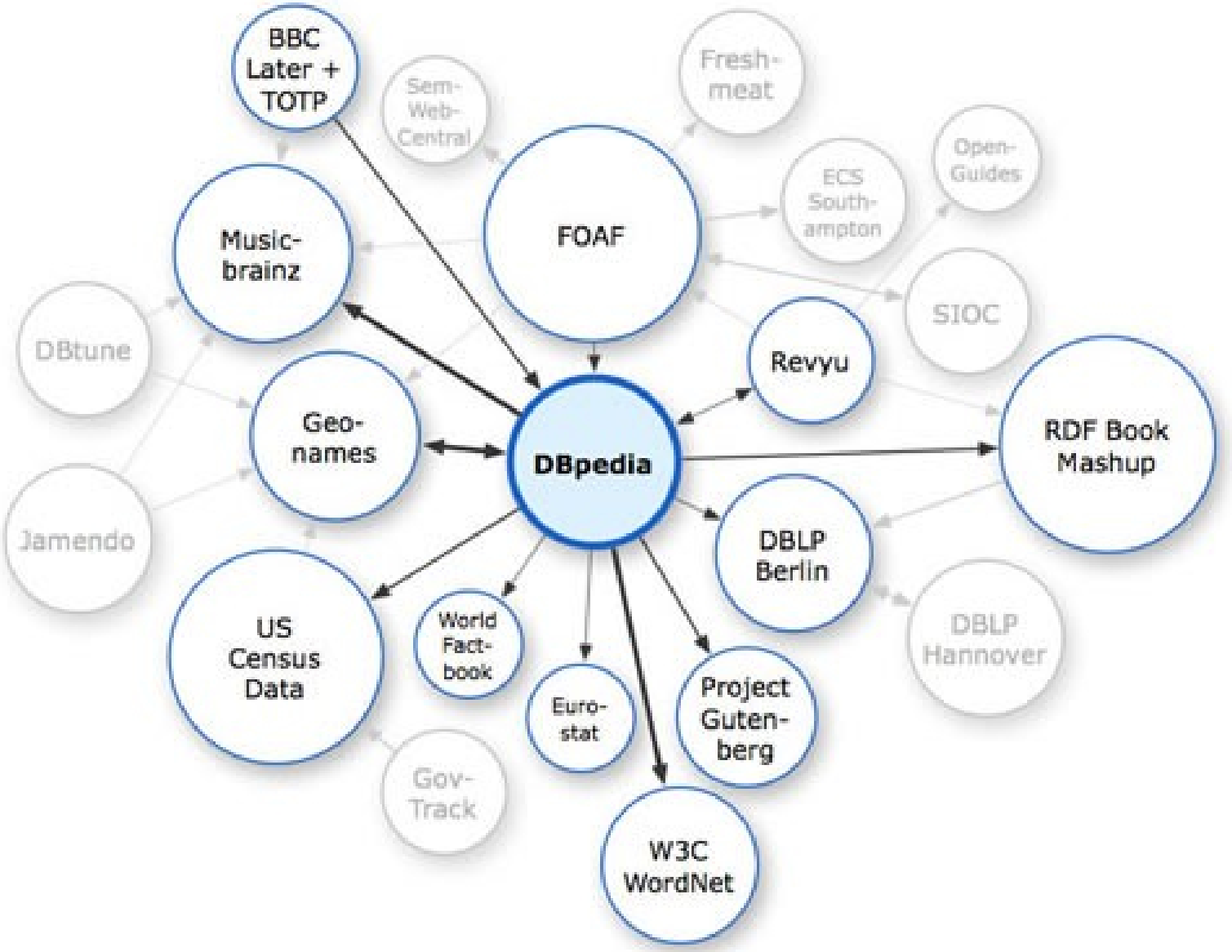


Figure 1. "is a" relation example

Projet sur le français : WOLF (Wordnet Libre du Français)

Traitement du langage : DBpedia

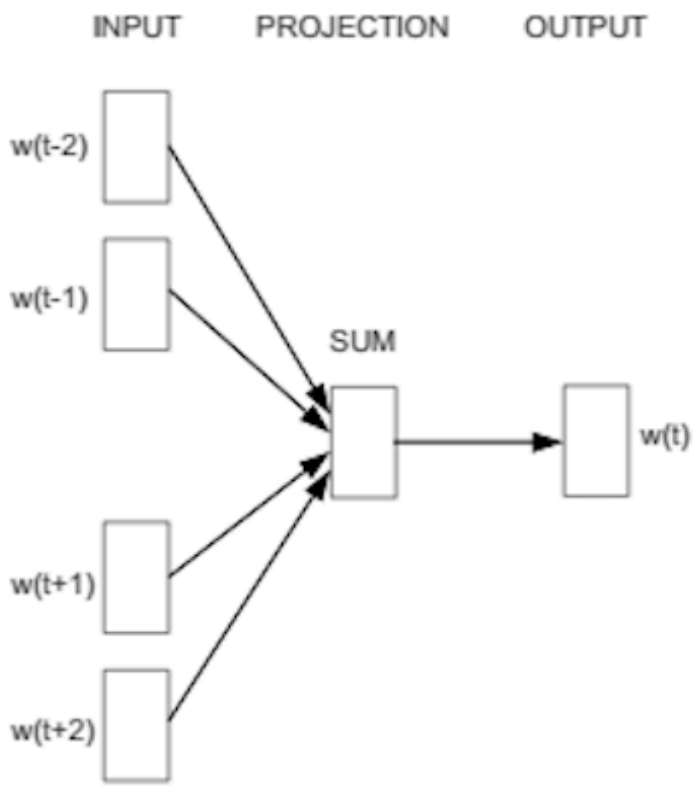


Traitement du langage : word embeddings

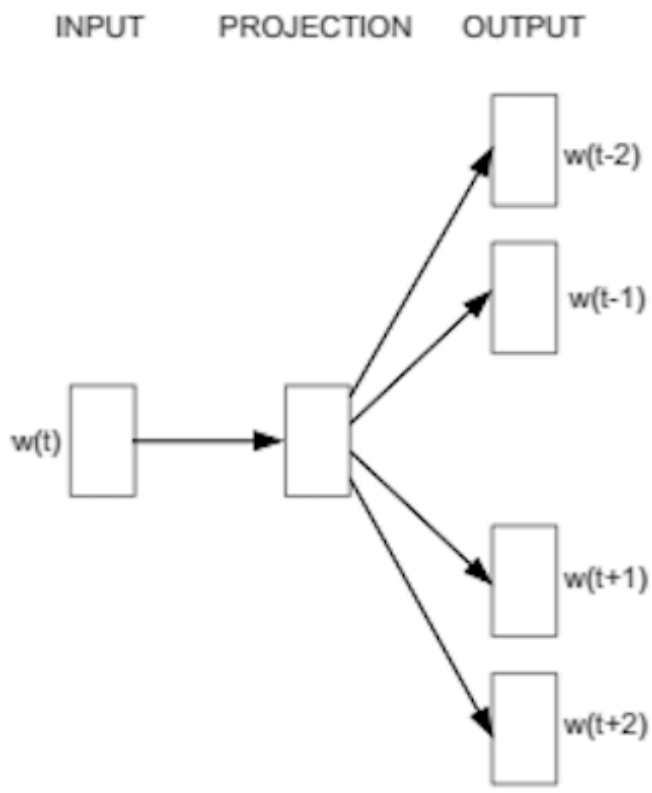
mot = indice dans un dictionnaire (dimension > 30000)
mot = vecteur “sémantique” (dimension < 1000)

- word2vec
- CBOW/Skip-Gram
- GloVe
- Thought vector (pour des phrases ou même des documents entiers)
- ...

Traitement du langage : word embeddings

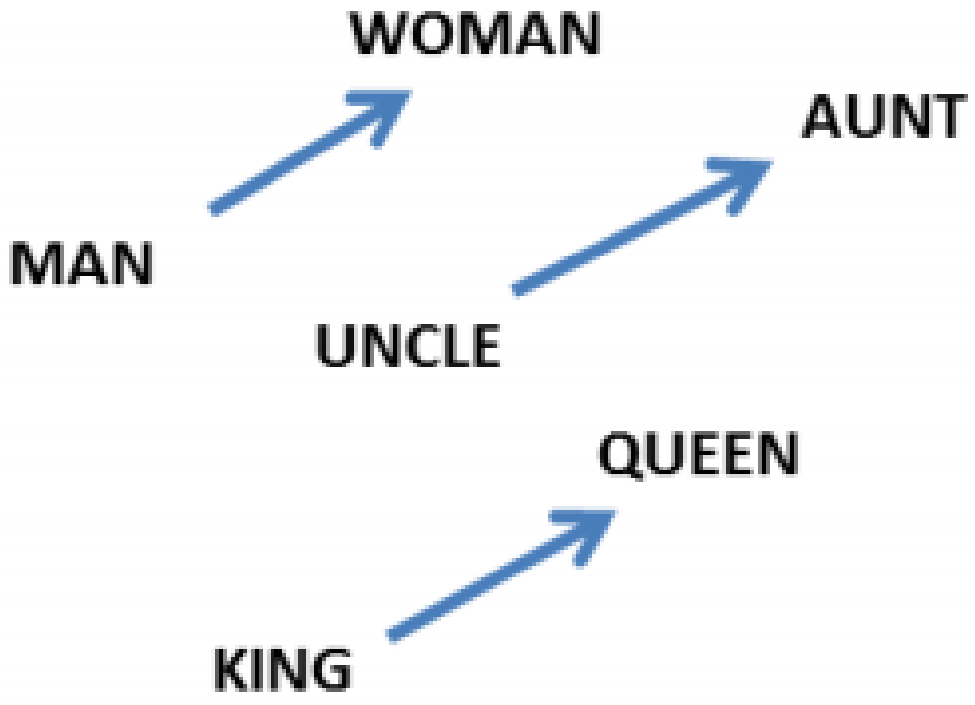


CBOW



Skip-gram

Traitement du langage : word embeddings



Démo dans l'espace word2vec

Visualisation de l'espace word2vec

Word Embeddings à télécharger