Вера Латышкина

SMT – Статистический машинный перевод – машинный перевод, основанный на статистических предсказаниях, а не только на правилах (как RBMT).

Мы хотим думать, что любую фразу на языке X можно декодировать и перекодировать на язык Y.

В такой модели предпочтение отдаётся наиболее правдоподобному переводу, то есть, наиболее вероятному соответствию единиц одного языка другому, который высчитывается статистически. То есть, чем чаще предложение X на входящем языке совпадёт по значению с предложением Y на принимающем в параллельном корпусе – тем правдоподобнее перевод.

Так, например, если в параллельных корпусах русского и английского языков у слова dog слово «собака» будет более частотным вариантом соответствия, чем слово «пёс», то «собака» будет более правдоподобным переводом, и система выберет его.

Чем больше выборка слов в параллельных корпусах (и чем чаще мы по ним проходимся через итерацию), тем выше вероятность (и точность) соответствий.

Компоненты модели SMT:

1. *N-граммная языковая модель*:

Нужна, чтобы моделировать *вероятностное распределение* конструкций на уровне слов или фраз в языке Y. Так мы можем искать вероятности продолжения токенов, т.е. искать/проверять лексическую/грамматическую сочетаемость («ein» -> «kleines»/«buch»)

1. *Модель перевода (t-model)* – матрица:  
     
   В ней собираем *статистику* соотвествий фраз в паралелльном корпусе, ищем переводческие соответствия X - Y и моделируем их с помощью теории *вероятности*:
   1. Допускаем, что любое предложение языка Y может быть "искаженной" версией некой фразы на языке X
   2. Ищем наиболее правдоподобные соответствия X – Y

В итоге получаем соответствия типа «my name is» -> «меня зовут» (вместо дословного и грамматичного «моё имя (есть)»), «ich schwimme gern» -> «я люблю плавать» (вместо «я плаваю охотно»)

1. *Декодер* – ищем наиболее грамматичные и лексически правдоподобные результаты, отбираем среди *гипотез* один результат