Статистические методы в машинном переводе: проблемы роста

Технологии перевода PROMT

- Rule-based перевод
- Статистический перевод
- Гибридный перевод

Статистические и гибридные системы перевода PROMT

- Разработка с 2009 года.
- Все языки rule-based системы PROMT, а также финский, китайский, японский, казахский, скандинавские языки.

Проблемы статистического и гибридного перевода

Проблемы данных:

- Источники
- Объем
- Предобработка

Проблемы перевода:

- Регистр
- Нестабильность перевода при использовании статистических компонентов

Статистическая/гибридная система перевода

Необходим:

- Параллельный корпус (для настройки модели перевода).
- Корпус на языке перевода (для настройки языковой модели).

Требуемый объем параллельных данных:

~ 1,5-2 млн. словоупотреблений (статистическая система).

~ 500 тыс. словоупотреблений (гибридная система).

Источники данных

Открытые источники:

- Параллельные и одноязычные корпусы в открытом доступе

- Интернет-ресурсы

Клиентские данные

Открытые источники

- Корпус Европарламента (20 языков, объем 10-50 млн. словоупотреблений)
- Корпус ООН (7 языков, объем 100-200 млн. Словоупотреблений)
- Субтитры (30 языков)
- Документация (PHP, OpenOffice, …)
- Новостные корпусы
- …

Подготовленные корпусы достаточно хорошего качества (txt/xml), но очень специфическая тематика.

- Интернет-ресурсы

Непараллельные, разнородные данные (html); много «мусора».

Настройка гибридной/статистической систем перевода

Настройка универсальной системы перевода

Использование
открытых корпусов
и Интернет-ресурсов

Настройка системы перевода для клиента

1. Можно настраивать систему только на клиентских данных

2. При малом объеме клиентского корпуса необходимо использовать дополнительные данные

- Общая большая модель перевода (клиентские + дополнительные данные).

- Две модели перевода (приоритетная на клиентских данных).

Предобработка данных

Клиентские данные тоже бывают «грязные»

Модуль предобработки данных PROMT:

1. Обработка данных в текстовом и xml-формате.
2. Валидация параллельных сегментов по соотношению длин.
3. Удаление дублирующихся сегментов.
4. Удаление непереведенных сегментов (модуль определения языка PROMT), сегментов с малым количеством алфавитных символов.
5. Разбиение сегментов на предложения (если в сегменте их несколько).
6. Нормализация знаков препинания, лигатур, html-сущностей.
7. Токенизация.
8. *Восстановление сегментов с неправильной кодировкой.*
9. *Проверка орфографии.*

Предобработка данных

Некоторые проблемы невозможно решить с помощью самых сложных алгоритмов прочистки.

Фрагмент параллельных англо-русских данных клиента (руководство пользователя):

The name
*“**Dinesh Kumar Garg**”*
will be written as:

Полное имя
*«**Иван Петрович Иванов**»*
будет записано следующим образом:

Проблемы перевода с использованием статистических компонентов

Проблема

восстановления регистра

Проблемы,

связанные с нестабильностью/ непредсказуемостью машинного перевода

Восстановление регистра при переводе

Существует стандартная модель восстановления регистра – фактически упрощенная статистическая модель перевода из нижнего регистра в комбинированный.

Восстановление регистра при переводе

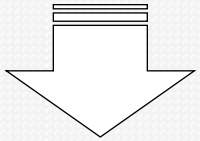
Гибридная система восстановления регистра PROMT:

1. Использование информации о регистре исходного текста (регистр первой буквы; все предложение в верхнем/нижнем регистре; сохранение регистра незнакомых слов).
2. Использование информации rule-based компонента (для гибридной системы перевода).
3. Статистическая модель восстановления регистра (большая универсальная модель + приоритетная модель на основе клиентских данных).

Нестабильность статистического компонента

Google:

Путин едет на желтой Калине



Putin goes to a yellow Mazda

Нестабильность статистического компонента

Исчезновение/появление лишних символов и пунктуационных знаков:

1.

Create a partition using the free space.
Создайте раздел, используя свободное пространство).

2.

Perform operations (e.g., ROP1, ROP2, and ROP3) on the data using a pattern.
Проведите операции (например, ROP1, ROP2 и ROP3,) на данных с помощью образца.

Решение:

1. Использование информации об исходном тексте.
2. Обработка невозможных сочетаний знаков.

Нестабильность статистического компонента

Проблемы со специальными конструкциями и именованными сущностями (имена, даты, адреса, номера телефонов и т.п.).

1.

Конгрессмен Джесси Л. Джексон-младший (демократ, Иллинойс) предложил множество поправок к закону.

Congressman Jesse Jr. L. Jackson (the democrat, Illinois) proposed a number of amendments to the law.

2.

If you want a refund, you must mail your written request to us at P.O. Box 45950, Omaha, NE 68145-0950.

Если Вы хотите компенсацию, Вам необходимо отправить свой письменный запрос по почте нам в почтовом ящике 45950, - Омаха, 68145-0950 штат Небраска.

Обработка специальных конструкций и именованных сущностей

Правильный перевод специальных конструкций и именованных сущностей – одно из основных требований клиентов.

Решение для гибридной системы перевода

• Использование информации rule-based компонента: специальные конструкции помечаются тегами и не обрабатываются статистическим компонентом.

Решение для статистических систем

• Использование парсеров PROMT (на стадии разработки).

Обработка специальных конструкций и именованных сущностей

Перевод специальных конструкций:

1.

Конгрессмен Джесси Л. Джексон-младший (демократ, Иллинойс) предложил множество поправок к закону.

- *Конгрессмен <NAME>Джесси Л. Джексон-младший</NAME> (демократ, Иллинойс) предложил множество поправок к закону.*
- *Congressman Jesse L. Jackson Jr.(the democrat, Illinois) proposed a number of amendments to the law.*

2.

If you want a refund, you must mail your written request to us at P.O. Box 45950, Omaha, NE 68145-0950.

- *If you want a refund, you must mail your written request to us at P.O. Box <ADDRESS>45950, Omaha, NE 68145-0950</ADDRESS>.*
- *Если Вы хотите компенсацию, Вам необходимо отправить свой письменный запрос по почте нам в почтовом ящике 45950, Омаха, штат Небраска 68145-0950.*

Обработка специальных конструкций и именованных сущностей

Обработка данных парсерами PROMT и замена специальных конструкций тегами на стадии обучения (исследования)

Существенное уменьшение объема словаря корпуса.

Улучшение качества выравнивания параллельных данных при обучении статистической/гибридной системы.

Спасибо за внимание!