Natural Language Processing

С развитием голосовых интерфейсов и чат-ботов, NLP стала одной из самых важных технологий искусственного интеллекта

Где применяется NLP

- поиск (письменный или устный);
- показ подходящей онлайн рекламы;
- автоматический (или при содействии) перевод;
- анализ настроений для задач маркетинга;
- распознавание речи и чат-боты,
- голосовые помощники (автоматизированная помощь покупателю, заказ товаров и услуг).

Глубокое обучение в NLP

Существенная часть технологий NLP работает благодаря глубокому обучению (deep learning) — области машинного обучения, которая начала набирать обороты только в начале этого десятилетия по следующим причинам:

- Накоплены большие объемы тренировочных данных;
- Разработаны вычислительные мощности: многоядерные CPU и GPU;
- Созданы новые модели и алгоритмы с расширенными возможностями и улучшенной производительностью, с гибким обучением на промежуточных представлениях;
- Появились обучающие методы с использованием контекста, новые методы регуляризации и оптимизации.

Этапы обработки

Сбор данных

Этапы обработки

Очистка данных
Ваша модель сможет стать лишь настолько хороша, насколько хороши ваши данные

Удалить все нерелевантные символы (например, любые символы, не относящиеся к цифро-буквенным).
Токенизировать текст, разделив его на индивидуальные слова.
Удалить нерелевантные слова — например, упоминания в Twitter или URL-ы.
Перевести все символы в нижний.

Этапы обработки

Выбор правильного представления данных

One-hot encoding

Классификация

Часто используется логистическая регрессия
Логистическая регрессия редсказывает вероятность возникновения события по значениям некоторых признаков

Инспектирование

Матрица ошибок

		Actual class		
		Cat	**Dog**	**Rabbit**
Predicted class	**Cat**	**5**	2	0
	Dog	3	**3**	2
	Rabbit	0	1	**11**

Учитывание структуры словаря

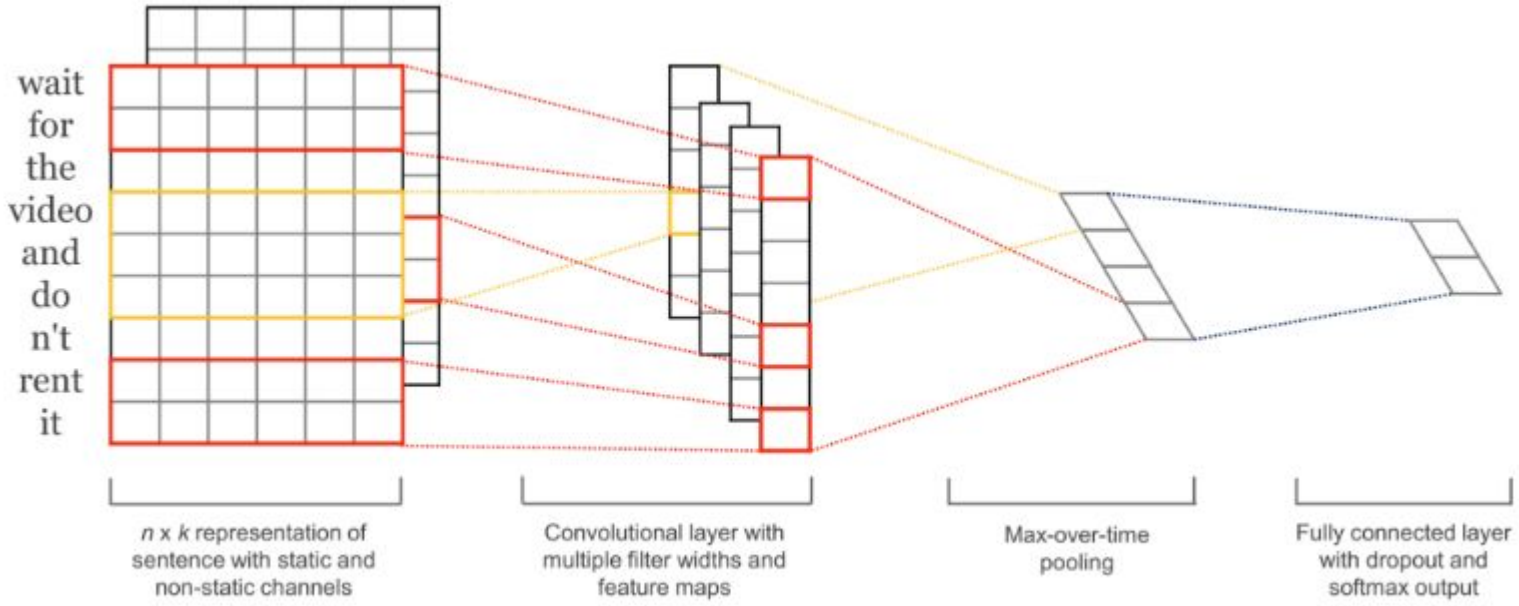
TF-IDF(*Term Frequency, Inverse Document Frequency*)

Применение семантики

Word2Vec
LIME
GloVe

Использование синтаксиса при применении end-to-end подходов

Glove
Word2Vec
CoVe



Машинный перевод

Используются статистика использования слов по соседству
Системы машинного перевода находят широкое коммерческое применение, так как переводы с языков мира — индустрия с объемом $40 миллиардов в год

- Google Translate переводит 100 миллиардов слов в день.
- Facebook использует машинный перевод для автоматического перевода текстов в постах и комментариях, чтобы разрушить языковые барьеры и позволить людям из разных частей света общаться друг с другом.

eBay использует технологии машинного перевода, чтобы сделать возможным трансграничную торговлю и соединить покупателей и продавцов из разных стран.
Microsoft применяют перевод на основе искусственного интеллекта к конечным пользователям и разработчикам на Android, iOS и Amazon Fire независимо от доступа в Интернет.

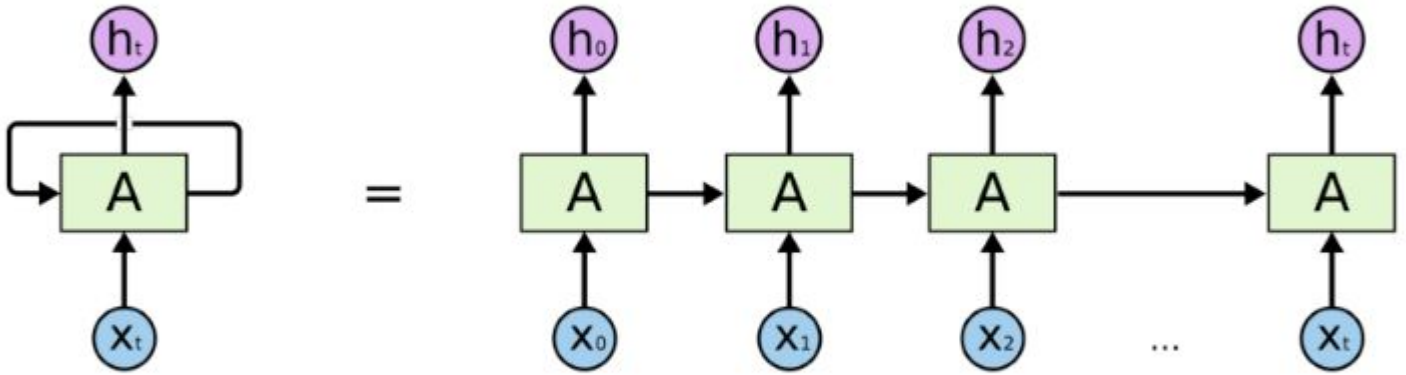
Systran стал первым поставщиком софта для запуска механизма нейронного машинного перевода на 30 языков в 2016 году.

Проблемы машинного перевода

Традиционные системы вынуждены использовать параллельный набор текстов для перевода.
До появления нейросетевого перевода, применялся статический подход для перевода, основанный на теореме Байеса.

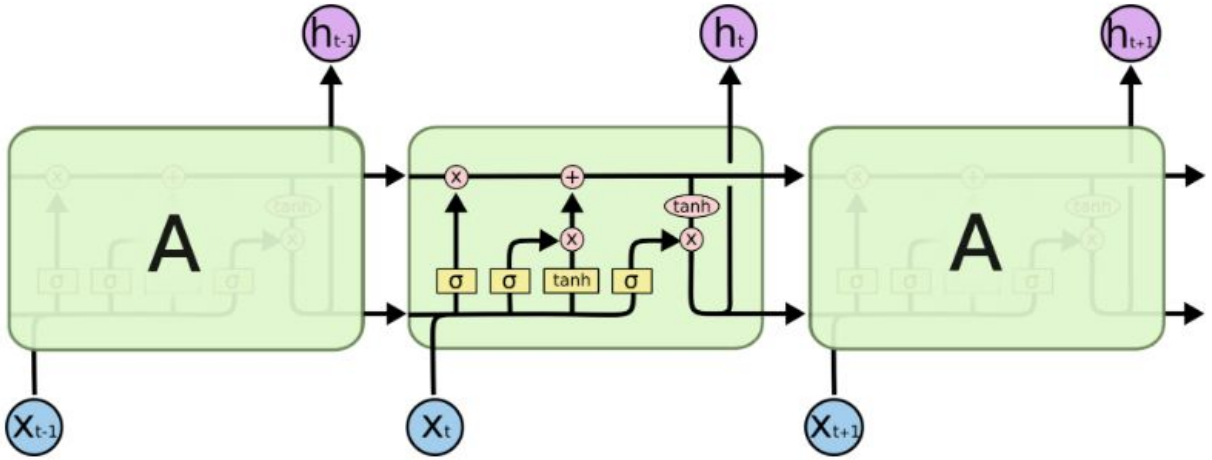
Нейросетевой машинный перевод

Используются реккурентные нейронные сети



Нейросетевой машинный перевод

Сети LSTM(Long short-term memory)
Позволяют обнаруживать долговременные зависимости



Нейросетевой машинный перевод

- Сквозное обучение: параметры в NMT (Neural Machine Translation) одновременно оптимизируются для минимизации функции потерь на выходе нейросети.
- Распределенные представления: NMT лучше использует схожести в словах и фразах.
- Лучшее исследование контекста: NMT работает лучше с контекстом, чтобы переводить точнее.
- Более беглое генерирование текста: перевод текста на основе глубокого обучения намного превосходит по качеству статический метод.
- Проблема исчезновения градиента
- LSTM решают данную проблему

Голосовые помощники

QA Системы

Идея QA систем заключается в извлечении информации непосредственно из документа, разговора, онлайн поиска или любого другого места, удовлетворяющего потребности пользователя.
Существует оптимизированная архитектура глубокого обучения.
Dynamic Memory Network.

Dynamic Memory Network.

Архитектура

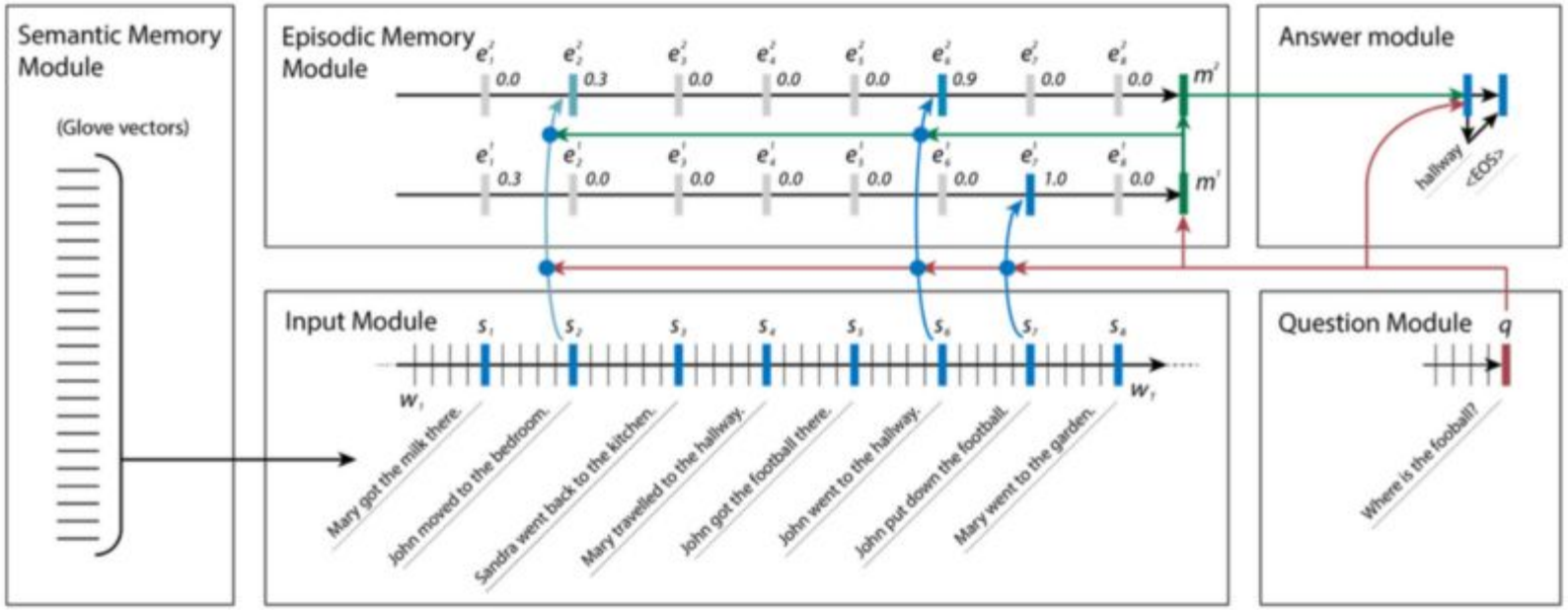


Illustration of DMN performing transitive inference.

Краткое изложение текста(Text Summarization)

Извлечение краткого содержания — важный инструмент для помощи в интерпретации текстовой информации.

Автоматическое извлечение краткого содержания

Схема работы:

1. Считается частота слов в полном тексте
2. N наиболее частых слов сохраняются
3. Каждое предложение оценивается по кол-ву частых слов
4. Первые M предложений сортируются с учетом положения в тексте

Сокращение текста

Извлекательный:
Извлекательный подход извлекает слова и фразы из оригинального текста для создания резюме
Примеры:
LexRank и TextRank

Сокращение текста

Абстрактный:
Абстрактный подход изучает внутреннее языковое представление, чтобы создать человекоподобное изложение, перефразируя оригинальный текст
Используется deep learning, тем самым данный подход достиг больших успехов
Примеры:
Facebook Neural Attention
Google Sequence-to-sequence
IBM Watson

- Конец