Beyond MT: Opening Doors for an NLP Pipeline

Alex Yanishevsky

Senior Manager, AI Deployments

AMTA VIRTUAL 2020

Overview

Primary Use Cases of MT

MT for NLP Pipeline

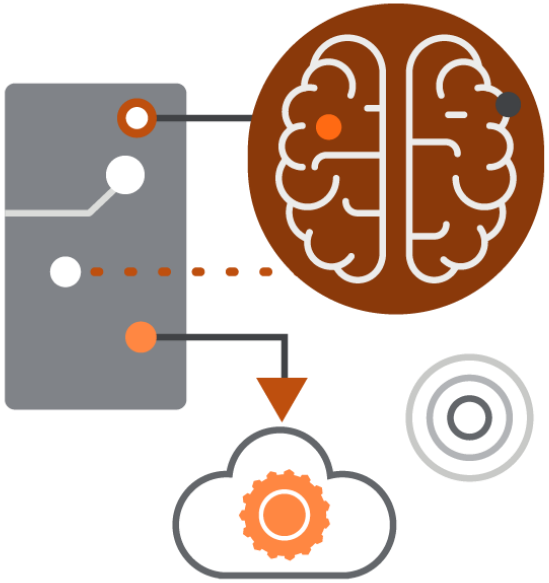
- Why?
- Before MT: Language identification
- After: MT Quality Estimation
- After MT: Social Listening
- After MT: Named Entity Recognition
- After MT: Dependency Parsing
- After MT: Keyword Search

Case Studies

Primary Use Cases of MT

Primary Use Cases of MT

- From and into English
- Generic or trained engines (domain, product, etc.)
- Informational (raw MT) including chat, forums, knowledge bases
- Post-editing (light, medium, full)
- Via MT connectors in TMS or CAT tools
- MT Quality Estimation



MT for NLP Pipeline

Why?

- ✓ **Many NLP packages (such as NTLK, Stanford CoreNLP or spaCy) not available or lag behind for non-English languages, e.g. readability for Flesch-Kincaid, POS tagging, dependency parsing, named entity recognition, stemming, lemmatization**
- ✓ **Insufficient data to train models**

Source: Memsource, AMTA 2020, Session C14

- Domains were defined using unsupervised machine learning on aggregate customer data, labels assigned manually
  - For non-English source languages, internal MT into English is applied first

NLP Pipeline

- Before MT: Language identification
- Machine Translation (generic or trained)
- After: MT Quality Estimation
- After MT: Social Listening
- After MT: Named Entity Recognition*
- After MT: Dependency Parsing
- After MT: Keywords

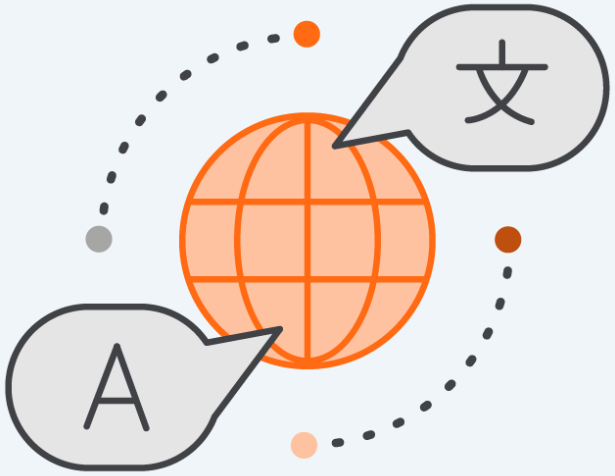
* Can also be done Before MT

Before MT:
Language Identification

For some domains such as litigation, a file or email may be multi-lingual. Thus, we need a way to identify the language(s) and pass them to MT in one request.

How to deal with this?

Language ID suite with **five** algorithms and majority polling Identification, MT and reassembly on a segment basis.



Example

Программное обеспечение защищено законодательством и международными соглашениями об авторском праве, а также законодательством и соглашениями о защите интеллектуальной собственности. Программное обеспечение не продается, а предоставляется в пользование по лицензии. Puede activar cierto software mediante una clave de licencia proporcionada por el servicio de soporte técnico de Luminex, enviando un mensaje a support@luminexcorp.com o llamando al 1-877-785-2323 o al 1-512-381-4397. 경기 부천에 있는 쿠팡 물류센터 관련 신종 코로나바이러스 감염증(코로나19) 환자가 급속도로 늘어나자, 정부는 내달 14일까지 수도권 내 모든 다중이용시설 운영을 한시적으로 중단하기로 했다. 다만, 수도권 내 초.중.고 등교 수업은 중지 없이 진행된다.

After MT: Quality Estimation

1. Readability
2. Adherence to style based on language models, edit distance, word embeddings
3. Segment length (word and character)
4. Complex words
5. Part of speech tagging
6. Build predictive models based on salient features

After MT: Social Listening

Brand Health

Evaluating public perception of brand and/or products.

Industry Insights

Analyzing discussions or hashtags related to specific industry.

Competitive Analysis

Analyzing competing brands or products.

Campaign Analysis and Event Monitoring

- Evaluating public perception of a campaign.
- Monitoring audience responses to a conferences and/or events.

After MT: Named Entity Recognition*

Recognition (Identification)
Deanonymization
Reassembly

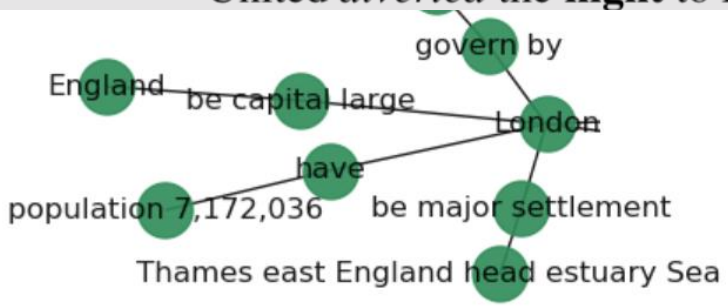
GDPR Compliance
HIPAA Compliance
Responsive (hot) document for litigation

* Can be done before MT

After MT: Dependency Parsing

1. **What is it?**
2. **How to do it? Dependency Parse Tree, Head-Dependent**
3. **Why do it?**

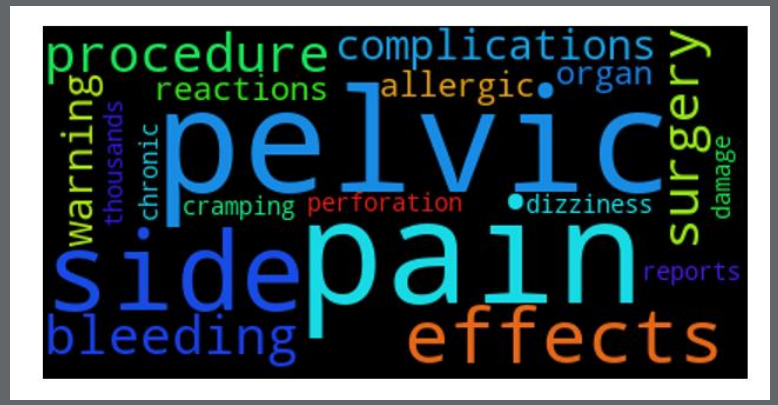
Relation	Examples with *head* and **dependent**
NSUBJ	**United** *canceled* the flight.
DOBJ	United *diverted* the **flight** to Reno.

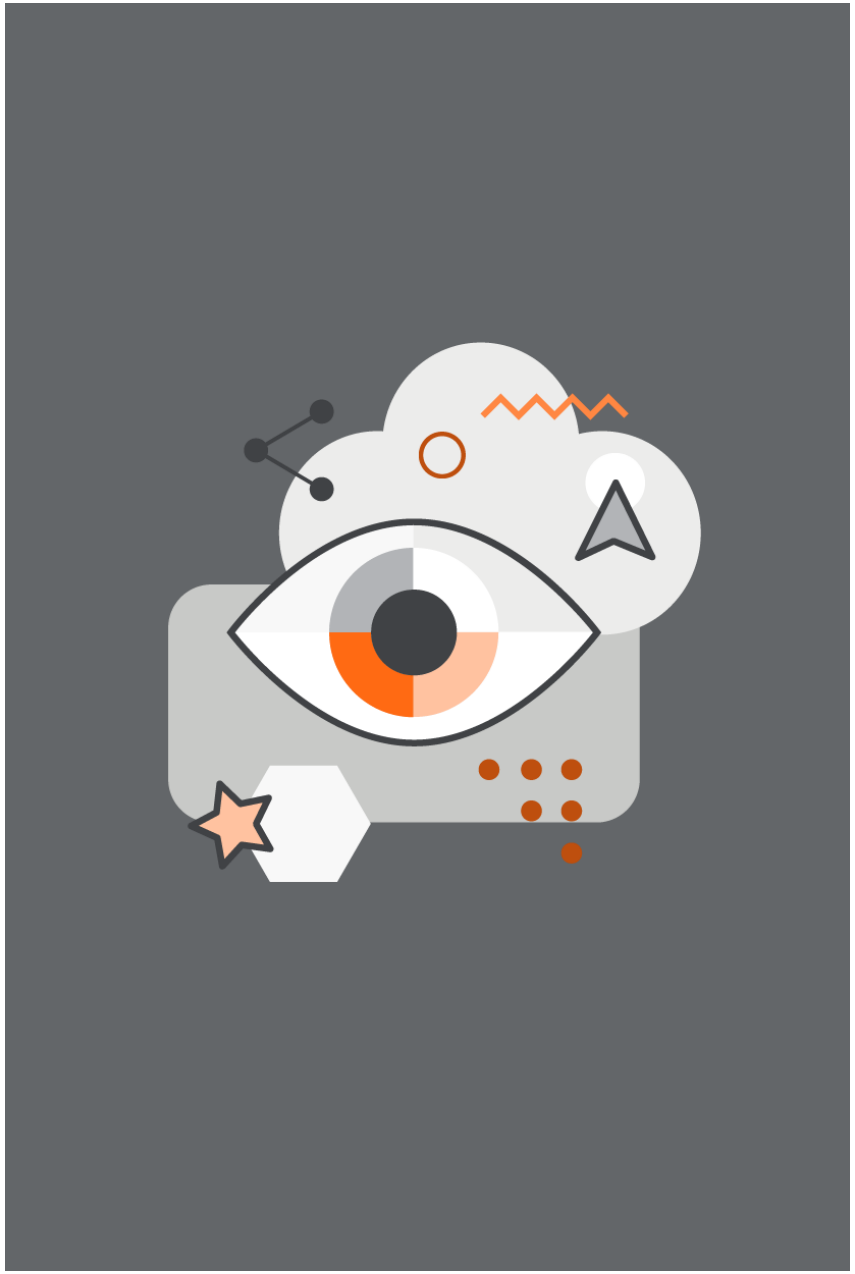


Source: https://medium.com/data-science-in-your-pocket/dependency-parsing-associated-algorithms-in-nlp-96d65dd95d3e

After MT: Keyword Search

- An example of a word cloud with salient terms for side effects of a drug





Case Studies

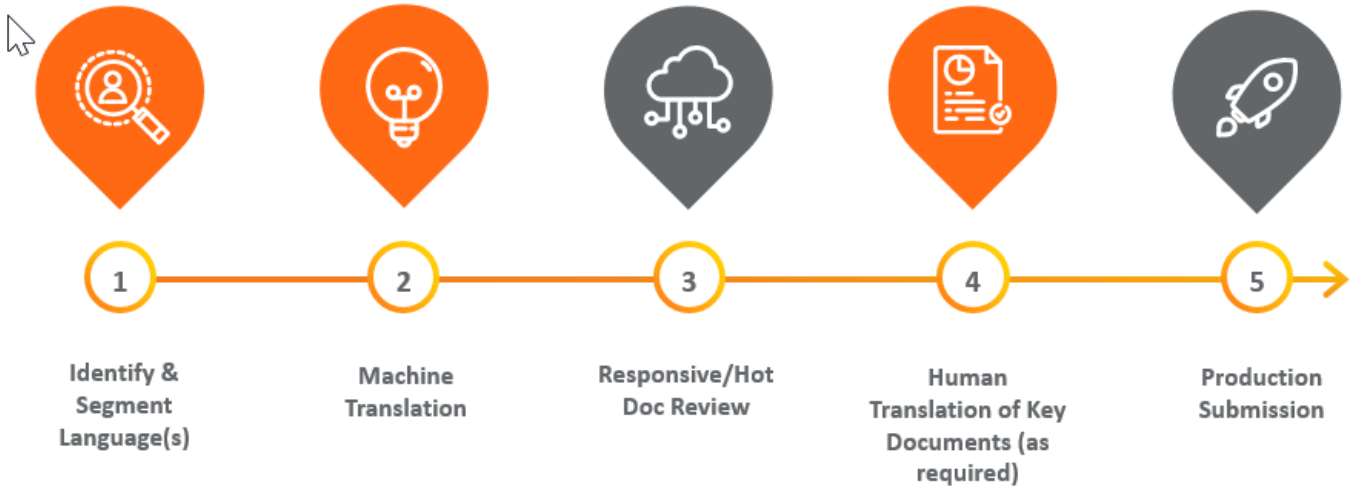
CASE STUDY

Litigation

Over 200 Million
words translated

Challenge

- Quick MT turnaround on 20K plus documents



Results

- Over 1 million USD saved versus human translation
- Saved over 2 months versus human translation
- Targeted selection of responsive documents

CASE STUDY

Life Sciences

Challenge

- Social listening for FR and ES
- Monitor responses of patients taking medication on social media channels

Solution

- Normalization of UGC
- Named Entity Recognition
- Customized sentiment analysis models including parsing ironic and sarcastic comments

Results

- Respond to patients' concerns
- Monitor and take action on adverse side effects
- Geographical, product and context distributions

Thank you

alexy@welocalize.com
https://www.linkedin.com/in/alexyanishevsky/

welocalize