



# Custom Translation Using KNIME and Keras

9<sup>th</sup> October 2018 | Mohammed Ayub and Joseph Gochal | Research and Data Analytics |  
National Fire Protection Association (NFPA)

# NFPA Introduction

**Vision - “ We are the leading global advocate for the elimination of death, injury, property and economic loss due to fire, electrical and related hazards”**



Non-profit organization delivering information and knowledge through more than 300+ Code and Standards relating to fire and electrical safety.

Providing training, education, research and outreach materials to more than 50,000 members around the world

Research and Data Analytics (RDA) group focusses on building interactive tools and services to enable fire data analytics



# Top Languages by Total Speakers

| Rank | Language                         | Total Speakers (millions) |
|------|----------------------------------|---------------------------|
| 1    | English                          | 1.12                      |
| 2    | Chinese (incl. Standard Chinese) | 1.1                       |
| 3    | Hindi/Urdu                       | 697                       |
| 4    | Spanish                          | 513                       |
| 5    | Arabic                           | 422                       |
| 6    | French                           | 285                       |
| 7    | Malay                            | 281                       |
| 8    | Russian                          | 264                       |
| 9    | Bengali                          | 262                       |
| 10   | Portugese                        | 236                       |



# NFPA Current State of Affairs

- Outsource most translation efforts, reviewing done by In-house experts
- Focused mainly on Spanish Portuguese and Arabic

## Code and Standards



- Colaborative programs like:
  - Latin America Chapters (6)
  - MENA Advisory Committee
  - Academic institutions in Sweden, UK

## Global Outreach



- ~80% Viewership and ~90 Product consumers are from North America Region. Rest is mostly spread across Asia and Middle East.
- NFPA web pages in Spanish

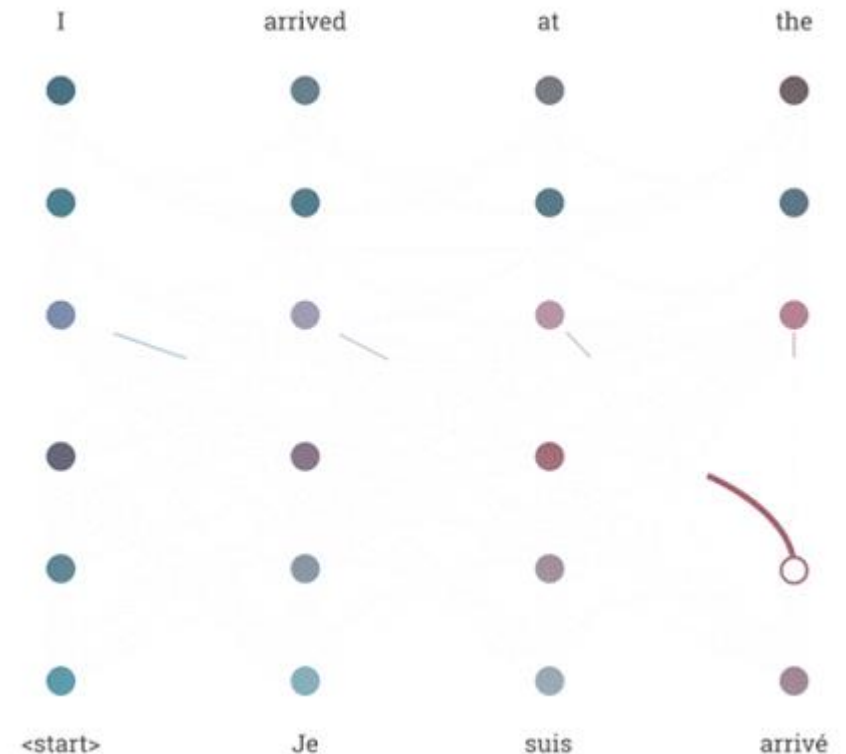
## Product Support



# Neural Machine Translation

- Converting a sequence of symbols in one language to another
- Encoder Decoder architecture
- Variations – RNN's, CNN's and mixed type

Decoding



Sample:

“Section 3.2.1.2 Fire-rated glazing assemblies marked as complying with hose stream requirements (H) shall be permitted in applications that do not require compliance with hose stream requirements.”

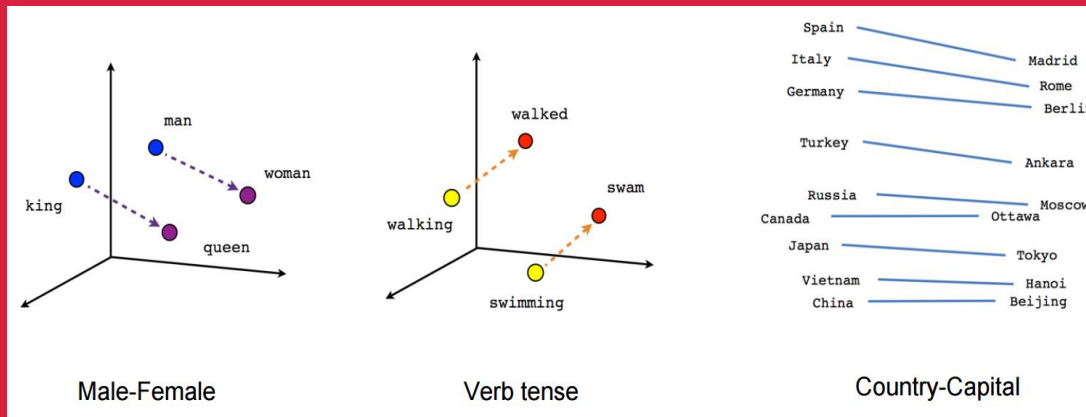
**Preprocess:**

“Arc-Fault Circuit Interrupter (AFCI)”  
“Circulating Closed-Loop Sprinkler System”

## Considerations

**Word Embedding:**

**Sub Words Units – BPE:**



Data Compression Algorithms for Word Segmentation

solar system (English)

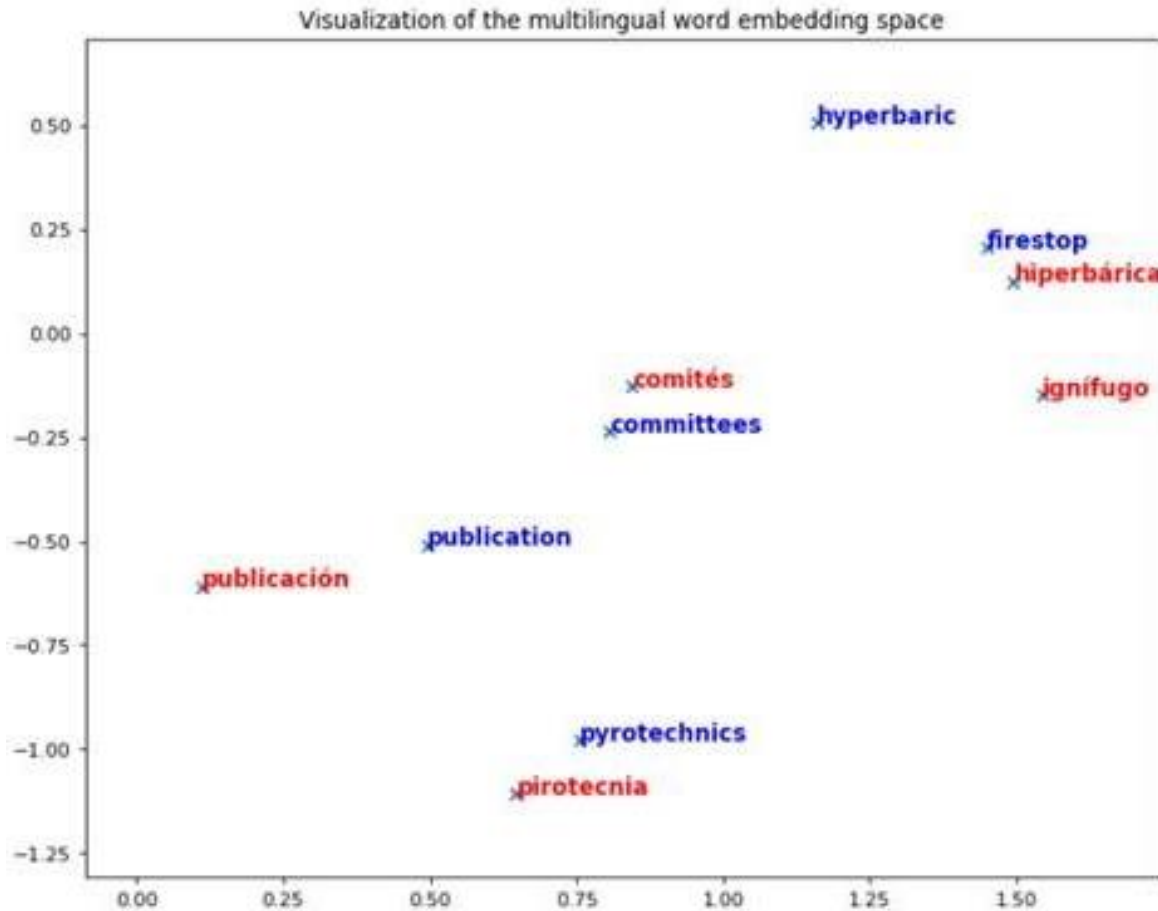
Sonnensystem (Sonne + System) (German)

Naprendszer (Nap + Rendszer) (Hungarian)

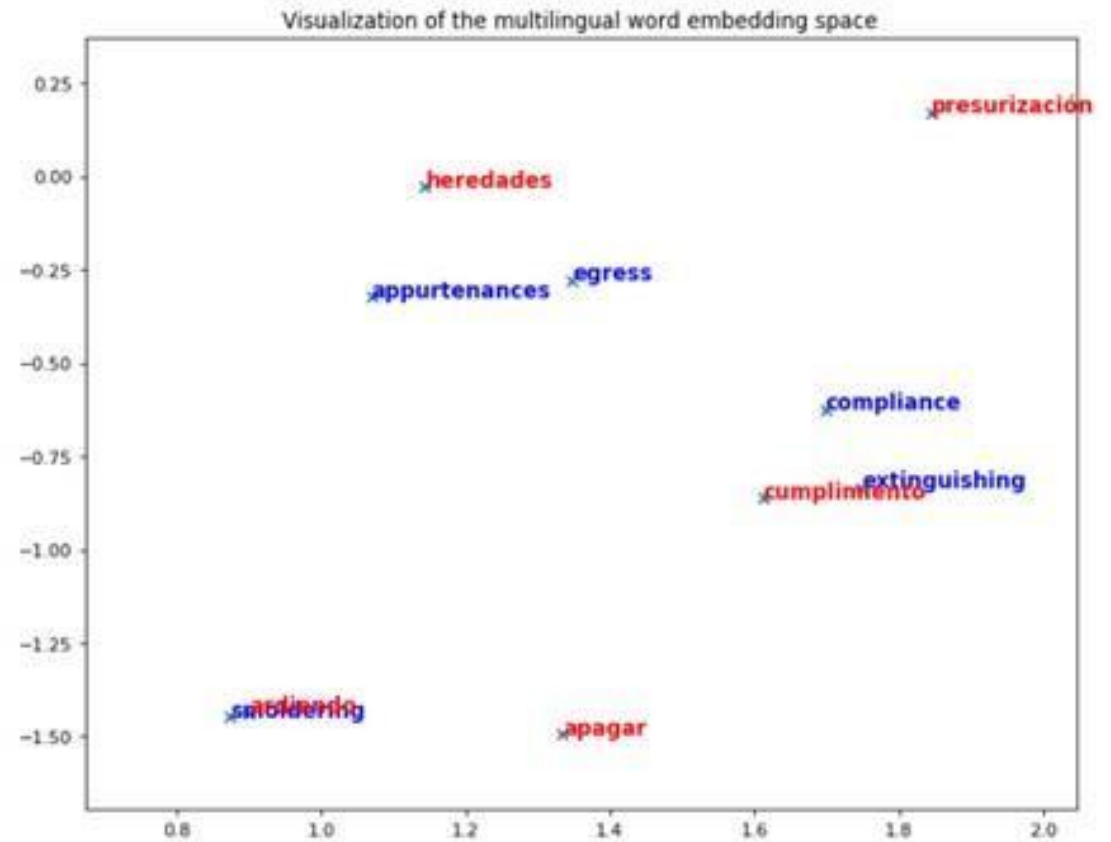


# Why Domain Adaption is hard ?

Good



Not so good



# Using KNIME for NMT

- Better integration, management, and documentation of workflows for internal and Cloud services
- Use KNIME Server to provide an easy user interface for parameter tuning
- Enable subject matter and language experts to review and edit NMT produced translations





# Keras-TensorFlow KNIME Pipeline



# Keras Pipeline (continued...)

English\_Spanish\_NFPA 2018-11-09 15.27.46

## Please enter the details for your Input Dataset Configuration

See below for other language data files to download

Training Data File

Change File

Selected file "spa.txt" (7 MB)

Output Folder

D:\Users\mayub\downloads\spa-eng\knime\_try2

No Of Rows (Filter)

5000

Traning Model Type

Basic Encoder Decoder  
Advanced Birectional LSTM  
Attention based LSTM

[Download more Bilingual Sentence Pairs](#)

Choose the Type of model you want to Run

< Back

Discard

Next >

# Keras Pipeline (continued...)

English\_Spanish\_NFPA 2018-11-09 15.48.36

## Model Hyperparameters: Attention Based

Sentence Max Length

10

BatchSize Training

64

Train Epochs

5

Train Optimizers

sgd  
adam  
adamax  
rmsprop

Train Loss Function

categorical\_crossentropy  
sparse\_categorical\_crossentropy  
mean\_absolute\_error  
mean\_squared\_error  
mean\_absolute\_percentage\_error  
mean\_squared\_logarithmic\_error

Model Output Directory

D:\Users\mayub\downloads\spa-eng\knime\_try

## Inference Details:

Choose Prediction File

Change File

Default file "english-spanish-test.pkl" (2 MB)

Back

Discard

Next

# Keras Pipeline (continued...)

## NFPA English to Spanish Translations

Custom Keras Model Results

Show  entries

Search:

| <input type="checkbox"/> | RowID | ↕ | source                    | ↕ | target              | ↕ | predicted | ↕ |
|--------------------------|-------|---|---------------------------|---|---------------------|---|-----------|---|
| <input type="checkbox"/> | Row0  |   | tomemonos algo            |   | lets have a drink   |   | me        |   |
| <input type="checkbox"/> | Row1  |   | cierra la escotilla       |   | close the hatch     |   | is it     |   |
| <input type="checkbox"/> | Row2  |   | empezad a cantar          |   | start singing       |   | i it      |   |
| <input type="checkbox"/> | Row3  |   | eso me sirve              |   | thatll do           |   | was       |   |
| <input type="checkbox"/> | Row4  |   | por que estas tan ocupado |   | why are you so busy |   | tom tom   |   |

Showing 1 to 5 of 1,000 entries

Previous **1** 2 3 4 5 ... 200 Next

### BLEU Scores for above Translations

Show  entries

Search:

| <input type="checkbox"/> |  | bleu_1               | ↕ | bleu_2                  | ↕ | bleu_3                  | ↕ | bleu_4                  | ↕ |
|--------------------------|--|----------------------|---|-------------------------|---|-------------------------|---|-------------------------|---|
| <input type="checkbox"/> |  | 0.021788918484587333 |   | 1.5441935317656938e-155 |   | 4.9448566696872296e-186 |   | 4.1108795915284684e-232 |   |

Showing 1 to 1 of 1 entries

Previous **1** Next



# AutoML KNIME Test Pipeline

## Google AutoML Translation Engine

This workflow accepts your GCP Credentials(json file) and test file containing Domain specific assests for translation (eg. Codes, journals, etc.)

### Setting up the environment

Setting up Environment



### Preparing the dataset

Display Datasets



### Training the model

Show Models



### Predicting/Inference

Predict Test Data



### Display Results

Display Results



# AutoML Pipeline (continued...)

google\_nmt\_WF 2018-11-09 15.38.15

Google Credentials File (JSON)

[Change File](#)

Selected file "My First Project-8bd0fb2d1072.json" (2 KB) \*

**ProjectName**

big-rig-221118

**Location**

us-central1

[← Back](#)

[Discard](#)

[Next >](#)

# AutoML Pipeline (continued...)

## Trained Models

Make Selection to test data

Show 10 entries

Search:

| ⊕ | RowID     | display_name             | dataset_id             | modelid                | create_time                                   | BLUE/Base BLUE  | Source/Target  | Sentences Pairs Evaluated | model_path                    |
|---|-----------|--------------------------|------------------------|------------------------|---|---|--|---------------------------|-------------------------------|
| ○ | Row0_Row0 | en_es_v1_v20181031193247 | TRL5376789144922460904 | TRL3188162022234012630 | seconds:<br>1541054668<br>nanos:<br>445609000 | bleu_score:<br>63.01748752593994<br>base_bleu_score:<br>43.05738806724548 | source_language_code:<br>"en"<br>target_language_code:<br>"es" | 3700                      | projects/3901<br>central1/mod |

Showing 1 to 1 of 1 entries

Previous 1 Next

Back

Discard

Next

## Test your Model on New Sentences

Please avoid passing huge files. Also check your daily API limits on the Google Console

Test File

Change File

Selected file "test\_file.txt" (2 KB)

Test File Output Directory:

D:\Projects\Jumpy Parrot- Neural Machine Translation\AutoML Test

Back

Discard

Next



# AutoML Pipeline (continued...)

Open for Innovation  
**KNIME** WebPortal

Settings Logout

google\_nmt\_WF 2018-11-09 01.05.16

Download Predicted Translations  
[Download Predicted Translations](#)

## NFPA English to Spanish Translations

Google AutoML Results

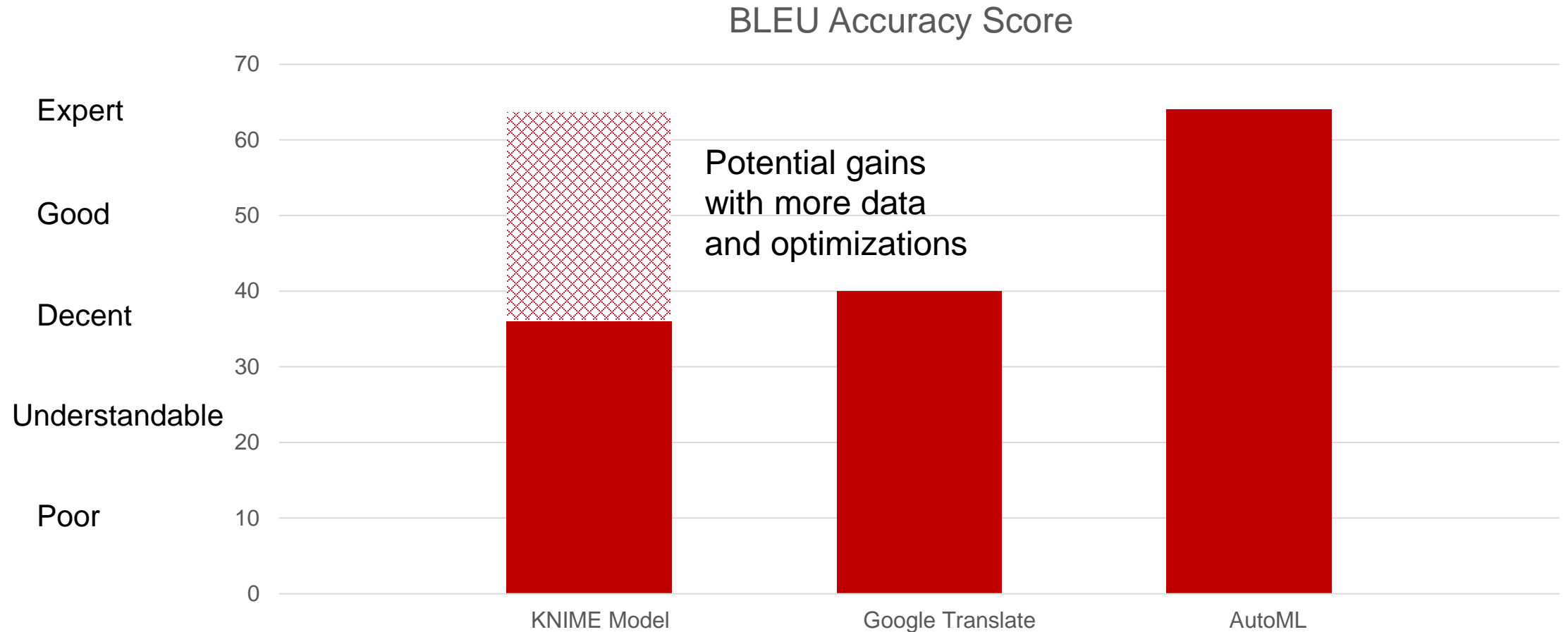
Show  entries  Search:

| <input type="checkbox"/> RowID <input type="text" value=""/> | English <input type="text" value=""/>   | Spanish <input type="text" value=""/>  |
|--|---|--|
| <input type="checkbox"/> Row0                                | 760.145 current-carrying continuous line-type fire detectors.   | 760.145 detectores de incendios de líneas continuas portadoras de corriente.   |
| <input type="checkbox"/> Row1                                | (f) in hoistways. in hoistways, power-limited fire alarm circuit conductors shall be installed in rigid metal conduit, rigid nonmetallic conduit, intermediate metal conduit, liquidtight flexible nonmetallic conduit, or electrical metallic tubing. for elevators or similar equipment, these conductors shall be permitted to be installed as provided in 620.21. | (f) en fosos de ascensores. en los fosos de los ascensores, los conductores de los circuitos de alarma de incendio de potencia limitada deben instalarse en conduit metálico rígido, conduit no metálico rígido, conduit metálico intermedio, conduit no metálico flexible hermético a los líquidos o tubería eléctrica metálica. para ascensores o equipos similares, debe permitirse que estos conductores se instalen como se establece en la sección 620.21. |
| <input type="checkbox"/> Row2                                | (g) other applications. for other applications, power-limited fire alarm circuit conductors shall be separated by at least 50 mm (2 in.) from conductors of any electric light, power, class 1, non-power-limited fire alarm, or medium-power network-powered broadband communications circuits unless one of the following conditions is met:                        | (g) otras aplicaciones. para otras aplicaciones, los conductores de circuito de alarma de incendio de potencia limitada deben estar separados como mínimo 50 mm (2 pulgadas) de los conductores de cualquier sistema de alumbrado eléctrico, de fuerza, de clase 1, de alarma de incendio de potencia no limitada o de red de potencia media. circuitos de comunicaciones de banda ancha energizados, a menos que se cumpla una de las siguientes condiciones:   |
| <input type="checkbox"/> Row3                                | (1) either (a) all of the electric light, power, class 1, non-power-limited fire alarm, and medium-power network-powered broadband communications circuit conductors or (b) all of the power-limited fire alarm circuit conductors are in a raceway or in metal-sheathed, metal-clad, nonmetallic-sheathed, or type uf cables.  | (1) ya sea (a) todos los conductores de los circuitos de alumbrado de potencia, de clase 1, de alarma de incendio de potencia no limitada y de comunicaciones de banda ancha energizados por una red de energía eléctrica, o (b) todos los circuitos de alarma de incendio de potencia limitada. los conductores están en una canalización o en cables con forro metálico, con blindaje metálico, con forro no metálico o cables tipo uf.                        |





# Results



# Results –

|              | BLEU (w/o training on NFPA Data) | BLEU (with training on NFPA Data) | Performance Gain (BLEU points) |
|--------------|----------------------------------|-----------------------------------|--------------------------------|
| Auto ML      | 43.1                             | 63                                | ~20                            |
| KNIME models | 12.5                             | 36.0                              | ~24                            |



# Results –

AutoML Custom:

para mayor información sobre los sistemas de alarma de incendio ver el artículo 760.

Official NFPA Translation

para mayor información sobre los sistemas de alarma de incendio, ver el artículo 760.

KNIME Model

para mayor información sobre los sistemas de alarma de incendio, véase el artículo 760.

for further information for fire alarm systems, see article 760.



# Example – Technical Terminology

## NFPA 1, section 12.7.4.2.1

Fire-rated glazing assemblies marked as complying with hose stream requirements (H) shall be permitted in applications that do not require compliance with hose stream requirements.

### AutoML:

Se deben permitir ensambles de vidrios resistentes al fuego marcados como que cumplan con los requisitos de la corriente de la manguera (H) en aplicaciones que no requieran cumplir con los requisitos de la corriente de la manguera.

### Official NFPA Translation

Deben permitirse conjuntos de montaje de vidrios certificados como resistentes al fuego señalizados, lo que indica que cumplen con los requisitos para chorros de manguera (H), en aplicaciones en las que no se requiera cumplir con los requisitos para chorros de manguera.

### KNIME Model

se debe permitir que los productos de vidrio, que cumplan con los requisitos de la corriente de flujo, se utilicen en aplicaciones que no requieran el cumplimiento de los requisitos de la corriente de flujo

shukran takk .  
kop-khuni merie  
efharisto spasiba danke  
kamsahamnida grazie  
mercici salamamat  
Thank You  
terima-kasih huala  
obrigado gracias  
dank-u  
keitos  
mahalo  
tak  
arigato