Esfuerzos recientes hacia la lectura de comprensión automática:

Una Revisión de Literatura

María Fernanda Mora Alba

Noviembre 18, 2016

Contenido

- Definición del problema
- Conjuntos de datos
- Modelos

Definición del problema

Contexto

- **Lectura de comprensión automática:** capacidad de un sistema para **leer** y **entender** textos en lenguaje natural a un nivel tal que es capaz de **responder preguntas**.
- **Enfoques tradicionales:** basados en reglas gramaticales y expertos -> caros y no escalables, progreso lento.
- **Ahora:** NLP, Extracción de información y Aprendizaje Máquina.

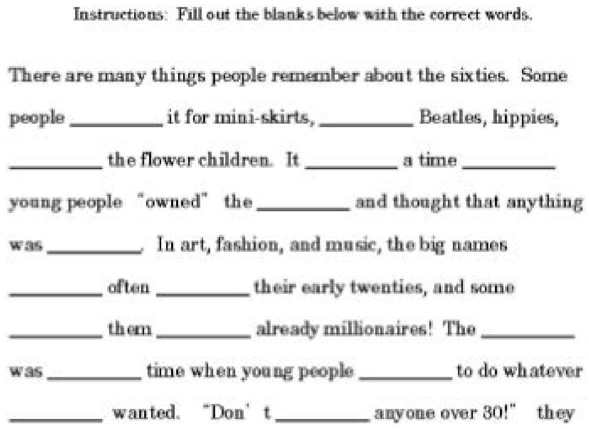
Problema

- Interés académico e industrial.
- Doble reto: conjuntos de datos y modelos.
- Doble problema:
    - Conjuntos de datos de calidad pero pequeños (human-annotated) vs grandes pero sintéticos (cloze-style).
    - Modelos dependen de los datos y son complejos.

Conjuntos de datos

Datasets: *cloze style*

- **Racional:** generar conjuntos enormes automáticamente.
- **Construcción:** Se remueven palabras de chunks de textos.
- **Objetivo de un modelo:** rellenar palabras faltantes.



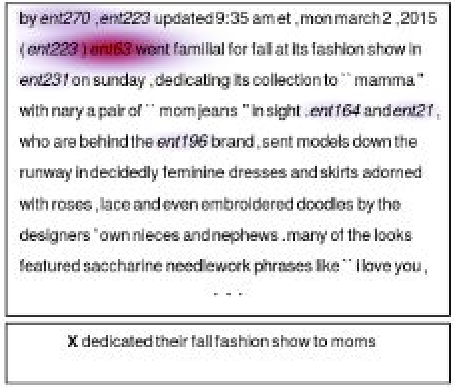
Instructions: Fill out the blanks below with the correct words.

There are many things people remember about the sixties. Some
people _________ it for mini-skirts, _________ Beatles, hippies,
_________ the flower children. It _________ a time _________
young people "owned" the _________ and thought that anything
was _________. In art, fashion, and music, the big names
_________ often _________ their early twenties, and some
_________ them _________ already millionaires! The _________
was _________ time when young people _________ to do whatever
_________ wanted. "Don' t _________ anyone over 30!" they

[Hadley and Naaykens 1999]

Datasets: *cloze style*

- **Hermann et al. 2015:** un millón de resúmenes de noticias de CNN y Daily Mail.
    - Anonimización de entidades



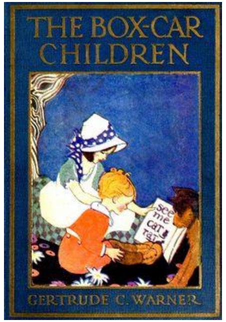
by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

[Hermann et al.2015]

- **Hill et al. 2015:** pasajes de libros para niños del proyecto Gutenberg.
    - (20 oraciones, 21va, palabra faltante)
    - Estructura narrativa



[Hill et al.2015]

Datasets: *cloze style*

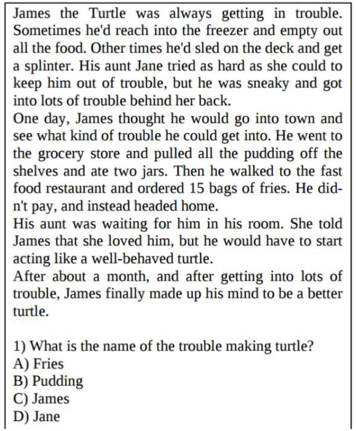
- Limitantes de datos tipo cloze-style:
  - No tienen preguntas tipo *factoid (W's)* que son fáciles de evaluar.
  - Poca inferencia de alto-nivel.
  - Algoritmos aprenden patrones de preguntas en vez de razonar sobre el significado.
- Cui et al. 2016: conjuntos similares en chino pero la evaluación hecha por humanos.

Datasets: *human annotated*

- Racional: generar conjuntos de datos de gran calidad, que realmente permitan evaluar la capacidad de entender textos
- Construcción: humanos escriben historias y/o preguntas sobre estas
- Objetivo del modelo: responder las preguntas
- Limitantes: escalabilidad, difícil usar modelos que requieren muchos datos

Datasets: *human annotated*

- **Richardson et al. 2013:** MCTest con 500 historias ficticias y 2K preguntas factoides por humanos.
  - Razonamiento causal, inferencia de relaciones, entendimiento de la historia
- **Tapaswi 2015:** MovieQA, 400 películas, 15K preguntas factoides de opción múltiple hechas por humanos.
  - Usa videos, subtítulos, scripts.



James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

[Richardson et al.2013]

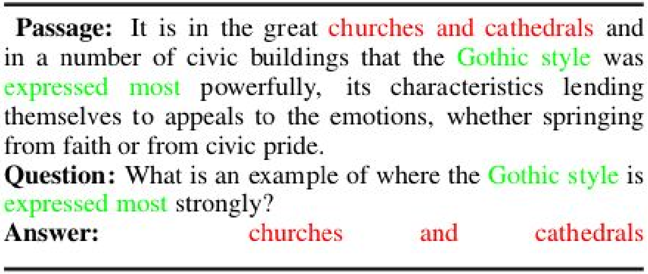
Q: Why does Forrest undertake a three-year marathon?
A: Because he is upset that Jenny left him

[Tapaswi et al.2015]

Datasets: SQuAD [Rajpurkar et al.2016]

Stanford Question and Answer Dataset

- +100K preguntas hechas por humanos sobre +500 artículos de Wikipedia
  - No hay respuestas candidato.
  - Las respuestas son sucesiones de tokens.
  - Preguntas y respuestas fueron hechas por humanos (usando crowdsourcing)



Passage: It is in the great churches and cathedrals and in a number of civic buildings that the Gothic style was expressed most powerfully, its characteristics lending themselves to appeals to the emotions, whether springing from faith or from civic pride.
Question: What is an example of where the Gothic style is expressed most strongly?
Answer: churches and cathedrals

Modelos

Modelos Tradicionales

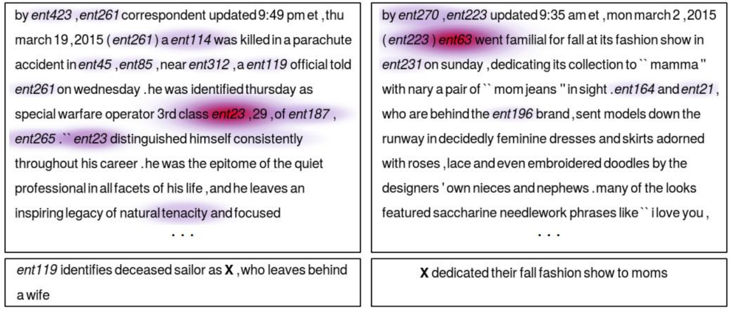
- Basados en reglas gramáticas hechas por expertos, anotación lingüística, parseo semántico, etc.
- Pipelines de submodelos que resuelven tareas específicas.
- Fallan al pasar de datos sintéticos a reales.
- No escalan.

Modelos Alternativos

- Requieren conjuntos enormes de datos (cloze-style).
- No requieren reglas ni expertos.
- Modelos end-to-end.
- Uso de redes neuronales (deep learning).

Modelos Alternativos

- **Hermann et al. 2015:**
    - RNN con mecanismos de atención para estimar *p(a | d, q)*
        - Lector LSTM profundo, atento e impaciente
    - Predicen un sólo token
- **Chen et al. 2016** hacen algo similar.



by *ent423* ,*ent261* correspondent updated 9:49 pm et ,thu march 19 ,2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* ,*ent85* ,near *ent312* ,a *ent119* official told *ent261* on wednesday .he was identified thursday as special warfare operator 3rd class *ent23* ,29 ,of *ent187* , *ent265* .`` *ent23* distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused

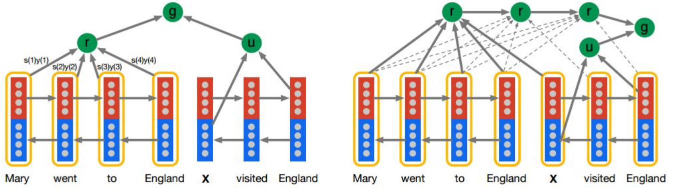
. . .

ent119 identifies deceased sailor as **X** ,who leaves behind a wife

by *ent270* ,*ent223* updated 9:35 am et ,mon march 2 ,2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday ,dedicating its collection to `` mamma " with nary a pair of `` mom jeans " in sight .*ent164* and *ent21* , who are behind the *ent196* brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers ' own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

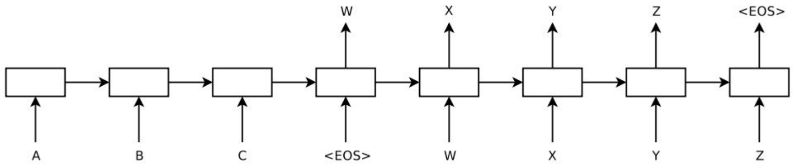


(a) Attentive Reader.

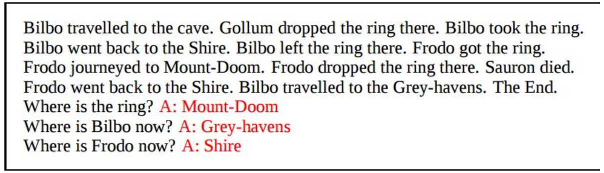
(b) Impatient Reader.

[Hermann et al.2015]

Modelos Alternativos



- **Yann et al. 2016:**
  - Modelos Sequence-to-sequence.
  - Flexibles: pueden generar múltiples tokens.
- **Weston et al. 2015, [Hill], [Sukhbaatar], [Kumar]**
  - Memory networks: *memorización*
  - Inferencia y memoria a largo plazo
  - Poco escalables



Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring.
Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring.
Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died.
Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End.
Where is the ring? A: Mount-Doom
Where is Bilbo now? A: Grey-havens
Where is Frodo now? A: Shire

Modelos Alternativos

- Kadlec et al.2016, Trischler et al. 2016:
    - Pointer Networks: Copiar tokens del párrafo como respuestas

Modelo baseline para SQuAD

- Rajpurkar et al. 2016:
    - Regresión logística con features hechos a la medida
    - 51% vs 87% del humano
    - Modelo sensible a:
        - Árboles de dependencia lexicalizados.
        - Tipos de respuestas: humano es más uniforme.
        - Divergencia sintáctica.

Modelo state-of-the art para SQuAD

- Wang et al. 2016
  - 2 modelos que usan match-LSTM (match una pregunta)
  - Pointer Networks (múltiples tokens)
  - 70.3% Score

Gracias