

Aprendizaje Activo para clasificación de preguntas sobre Datos Enlazados (Linked Data)

Teruel Milagro

Facultad de Matemática, Astronomía y Física

Universidad Nacional de Córdoba

Directores

Laura Alonso Alemany

Franco Luque

Índice general

1. Introducción	1
2. Definición formal del problema	4
2.1. Datos Enlazados y Sistemas de Respuesta	4
2.2. Quepy	7
2.2.1. Construcción de las consultas	8
2.2.2. Plantillas y sus expresiones regulares	10
2.3. Formalización del problema	11
2.4. Solución propuesta	12
2.4.1. De la Clasificación Supervisada al Aprendizaje Activo	13
2.4.2. Aprendizaje activo sobre características	14
2.4.3. Dualist	15
3. Representación de las preguntas	16
3.1. Subpatrones	16
3.2. Nombres y tipos de entidades	17
4. Arquitectura del sistema	19
4.1. Framework para aprendizaje automático sobre instancias y características	19
4.1.1. Funcionalidad	19
4.1.2. Parámetros	19
4.1.3. Selección de instancias	20
4.1.4. Selección de características	20
4.1.4.1. Information Gain	20
4.1.5. Maximización de la esperanza	21

5. Entorno de experimentación	22
5.1. Ejemplos seleccionados	22
5.1.1. Proceso previo	22
5.2. Usuarios Emulados	23
5.3. Experimentos realizados	23
5.3.1. Métricas utilizadas	23
5.3.2. Baseline	24
5.3.3. Hipótesis 1	25
5.3.4. Hipótesis 2	25
5.3.5. Experimento 3	25
5.3.6. Experimento 4	25
5.3.7. Experimento 5	26
Bibliography	27

Índice de figuras

2.1. Arquitectura de un sistema de aprendizaje activo	14
---	----

Capítulo 1

Introducción

Los sistemas de respuesta a preguntas son un área naciente del procesamiento del lenguaje natural y particularmente del área de recuperación de información.

Gupta and Gupta [2012] destacan que existen dos formas principales de buscar la respuesta a una pregunta de un usuario. La primera de ellas consiste de encontrar similitudes semánticas o sintácticas entre la pregunta y documentos de texto que pueden contener evidencias para la respuesta. La segunda, que abordaremos durante este trabajo, traduce la pregunta a un lenguaje formal para luego realizar consultas a una base de datos.

La reciente posibilidad de manejo de grandes volúmenes de datos ha permitido la formación de grandes bases de conocimiento públicas y disponibles online. Estas web semánticas u ontologías cambian ampliamente el paradigma utilizado hasta el momento, ya que estructuran los datos y permiten entonces extraer relaciones complejas entre sus entidades, como plantea Ou and Zhu [??].

Sin embargo, el primer paso para la resolución de una pregunta es la formalización de la misma a partir del texto ingresado por el usuario, independientemente del método de extracción de información empleado a continuación. Aún así, la mayoría de los sistemas se centran en la búsqueda de la respuesta más que en la correcta interpretación de la pregunta, y en general se limitan a textos cortos y preguntas puntuales.

Una aproximación simple a este problema es la de Quepy, un framework de traducción automática de preguntas en lenguaje natural a un lenguaje de consultas formalizado. El programador define una serie de plantillas para cada tipo de pregunta que el sistema pueda procesar y su correspondiente interpretación en la base de conocimiento elegida.

Aunque Quepy está diseñado para simplificar la tarea de construcción de dichas reglas, el trabajo necesario para lograr cobertura amplia es todavía prohibitivo por

varios motivos:

- Las plantillas deben ser desarrolladas por un experto de forma individual.
- El poder expresivo de las preguntas que soporta el sistema es lineal con respecto a la cantidad de plantillas generadas.
- Existe redundancia de información. Por ejemplo, para las preguntas “Who are the presidents of Argentina?” y “Who are the children of the presidents of Argentina?” se necesitan dos plantillas que contienen la misma información para resolver “presidents of X”.
- Existen numerosas preguntas que son equivalentes y que no necesariamente se representan con la misma plantilla. Por ejemplo las preguntas “Where is Angelina Jolie from?” y “Where was Angelina Jolie born?” tienen esencialmente la misma semántica.
- Debido a las grandes variaciones del lenguaje natural, se requiere un anotador experto para lograr una cobertura completa de todas las reformulaciones para una misma semántica.

De todas las dificultades anteriores nos enfocaremos en las dos últimas ya que las consideramos prioritarias y, al solucionarlas, podemos ampliar la cobertura de los sistemas construidos sobre Quepy significativamente. Nuestra propuesta es aplicar un clasificador automático sobre las preguntas donde cada clase es una interpretación de Quepy. De esta forma, podemos ligar muchas más reformulaciones de la misma pregunta a su correspondiente semántica y lograr mayor versatilidad.

La originalidad de nuestra aplicación se basa en utilizar como características las concordancias parciales con las plantillas de Quepy predefinidas por un programador. Consideramos que identifican claramente los aspectos relevantes que indican la correcta interpretación de la pregunta, y como tal son mejores representaciones.

Para evaluar nuestro sistema consideramos que comenzar con pocos patrones predefinidos nos ayudaría a percibir con más exactitud qué mejora podría generar el clasificador en Quepy. Por ello, y debido a que no existen grandes corpus etiquetados para reformulaciones de preguntas, planteamos que un enfoque de aprendizaje activo es lo más adecuado. El aprendizaje activo, como describe Settles [2009], permite entrenar el aprendedor con menor cantidad de instancias y es beneficioso cuando se cuenta con muchos ejemplos no etiquetados pero donde la mayoría no son relevantes. En nuestro entorno en particular se da este fenómeno, debido a que en un corpus

no anotado estándar pocas de las preguntas caerán dentro de alguna de las clases semánticas de los patrones iniciales.

Un enfoque novedoso que combina todos los conceptos anteriores es el de Settles [2011] en Dualist. Esta herramienta optimiza el aprendizaje activo no solo preguntado al usuario sobre instancias sino también sobre características de las mismas que las asocian a una clase. Junto con este desarrollo también incluye una serie de investigaciones sobre el rendimiento de tareas de clasificación con usuarios reales y simulados. Es por ello que tomamos como base este trabajo y lo adaptamos con una nueva implementación a nuestro problema.

Capítulo 2

Definición formal del problema

Tanto el problema que planteamos abordar como la solución propuesta son complejos de definir, ya que incluye numerosos conceptos del procesamiento de lenguaje natural. En la primera parte de esta sección definiremos un marco teórico para cada aspecto no central del problema. A partir de esta base, en la segunda parte daremos una definición propiamente dicha, seguida por una formalización de la solución.

2.1. Datos Enlazados y Sistemas de Respuesta

La cantidad de información disponible en internet es abrumadora, y sin embargo, aún no puede utilizarse en conjunto para extracción de información satisfactoriamente. Berners-Lee [2014] explican que este fenómeno se debe a que fragmentos de información que se refieren al mismo objeto o suceso no están relacionados entre sí

Christian Bizer and Berners-Lee [2009] definen los datos enlazados como información que cumple las siguientes características:

1. Puede ser leída automáticamente por una computadora.
2. Su significado está explícitamente definido.
3. Está conectada a fuentes de datos externas.
4. Puede ser conectada desde fuentes de datos externas a su vez.

Sin embargo, no existe un consenso o una definición formal sobre el tema. Berners-Lee [2014] describe en su artículo un protocolo orientativo para publicar datos enlazados en la web de tal forma que pudiera formarse una base de conocimiento global. Con el tiempo estas reglas se han tomado como un estándar para la construcción de ontologías, y en la actualidad existen espacios de información que contienen millardos de aserciones del mundo real.

Los datos enlazados se representan comunmente como una colección de tripletas siguiendo un lenguaje de descripción como RDF, tal como lo describe Brickley and Guha [2014]. Cada tripleta se compone de un sujeto, un predicado y un objeto, donde el predicado representa una relación entre el sujeto y el objeto. De esta forma se puede representar cualquier tipo de asociación entre entidades sin importar su complejidad, contruyéndolo a partir de relaciones parciales. El resultado es información organizada en forma de grafo donde cada nodo es una entidad y cada arista es una relación entre dichas entidades.

Las web semánticas u ontologías más populares en el momento son FreeBase ¹ y DBPedia ², aunque existen numerosos proyectos con dominios más acotados como WordNet ³. Estas plataformas son abiertas con interfaces fáciles de utilizar que permiten agregar nuevos datos, y como resultado se observa un rápido crecimiento en la cantidad de información disponible.

Estos sitios cuentan con puertos de accesos donde los usuarios pueden enviar consultas utilizando algún lenguaje formal. Aunque este servicio es accesible para cualquier persona, se requiere cierto nivel de conocimiento técnico para generar dichas consultas. Para dar acceso real a las masas a esta gran cantidad de información de requieren interfaces capaces de extraer datos a partir de consultas en lenguaje natural, es decir, sistemas de respuestas a preguntas.

Paralelamente, los sistemas de respuesta a preguntas pueden obtener grandes beneficios de una ontología. En lugar de buscar documentos o pasajes que puedan contener una respuesta, los datos enlazados pueden brindar información exacta. Además de ello, resulta más fácil procesar preguntas donde es muy poco probable encontrar la respuesta en un solo documento, por ejemplo, “¿Qué famosas actrices nacieron en el mismo país que Naima Akef?”. Desde los años 70 este tipo de software ha utilizado bancos de conocimiento estructurada que inicialmente eran bases de datos locales. Sin embargo, los resultados obtenidos no se destacaron particularmente. Con el desarrollo de las nuevas web semánticas la atención ha vuelto nuevamente hacia los datos relacionados.

Extraer información de una ontología no es difícil, sin embargo, como describe Unger et al. [2014], indentificar el sector de datos relevante a una consulta en lenguaje natural es un gran desafío. Se requiere para esto traducir el texto ingresado por el usuario en una consulta formal que pueda ser procesada por un motor de búsqueda

¹www.freebase.com

²www.dbpedia.org

³www.wordnet.princeton.edu

tradicional sobre datos enlazados. Una vez que se ha obtenido la información de la base, otra etapa de procesamiento convierte estos datos del formato legible por una computadora a un formato legible por el usuario. A continuación ilustramos con un ejemplo estas etapas utilizando una consulta en lenguaje MQL sobre la estructura de FreeBase.

Ejemplo 1.

1. Obtención de la pregunta.

What **is** the capital city of Argentina?

2. Generación de la consulta MQL.

```
{
  "type": "/location/country",
  "id": "/en/argentina",
  "capital": null
}
```

3. Obtención de la información.

```
{
  "result": {
    "capital": "Buenos_Aires",
    "type": "/location/country",
    "id": "/en/argentina"
  }
}
```

4. Generación de la respuesta en lenguaje natural.

The capital city of Argentina **is** Buenos Aires.

El primer desafío es identificar la entidad a la que se hace referencia en la pregunta, en nuestro caso, “Argentina”. Esta tarea se complicaría con nombre más complejos como “People’s Republic of China”. Las complicaciones de este etilo está ligadas a los sistemas externos de parseo y asignación de etiquetas morfosintácticas. Sin un buen procesamiento del lenguaje natural poco puede contruirse.

Adicionalmente, las consultas contienen no sólo información brindada por la pregunta del usuario, sino también datos asociados a la estructura de la base. Si en lugar de “/location/country” hubieramos utilizado “/location/location” la consulta hubiera devuelto un error, a pesar de que Argentina es también de tipo “/location/location”.

Unger et al. [2014] menciona también otros problemas que frecuentemente enfrentan este tipo de sistemas.

En las consultas utilizando MQL se detalla la estructura de la información y se completan los datos necesarios para identificar el objeto en la base de datos. Para obtener información sobre la entidad se nombran sus atributos, pero se les da un valor de *null*. El motor de búsqueda identifica estos campos y completa la información faltante. Este lenguaje es muy intuitivo y fue diseñado para ser accesible, pero no todos los lenguajes de consulta son tan simples como MQL.

Ejemplo 2. Consulta en SPARQL para la pregunta “How many episodes does Seinfeld have?”

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x1 WHERE {
    ?x0    rdf:type                                dbpedia-owl:TelevisionShow.
    ?x0    dbpprop:showName                        "Seinfeld"@en.
    ?x0    dbpedia-owl:numberOfEpisodes           ?x1.
}
```

La cantidad de información necesaria para construir esta consulta es mucho mayor mientras que su estructura no es simple de comprender. Sin embargo, pone en relevancia el uso de tripletas para representar la relación entre distintos nodos. En particular, la variable *?x1* representa el resultado, mientras que la variable *?x0* representa a la entidad de nombre “Seinfeld” y tipo “TelevisionShow”.

Cómo cierro esta sección? Quiero dar una buena introducción a las consultas para después explicar bien cómo funciona Quepy.

2.2. Quepy

Como se mencionó anteriormente, Quepy es un marco de trabajo para crear aplicaciones de respuesta a preguntas. Su objetivo principal es brindar una herramienta fácilmente adaptable a distintos dominios y distintos lenguajes de consultas. Los lenguajes soportados hasta el momento son MQL y SPARQL; ambos permiten consultas posteriores a FreeBase y DBPedia. Haremos un breve resumen a continuación sobre la arquitectura general de Quepy y sus principales características.

Una aplicación creada en Quepy tiene tres secciones principales:

Settings La configuración de Quepy incluye las herramientas de análisis sintáctico a utilizar, la URL del servidor para enviar las consultas, etc.

Templates Contiene las plantillas definidas por el creador de la aplicación. Cada plantilla es una expresión regular que combina distintos tipos de características como etiquetas POS y lemmas, lo que permite al sistema identificar la semántica de la pregunta únicamente en base a su sintáxis. Junto con la expresión regular, cada plantilla tiene una función de interpretación que toma las secciones de la pregunta que considera relevantes y las utiliza para construir una representación interna de la pregunta llamada Expresión.

DSL Son las siglas correspondientes a Lenguaje de Dominio Específico en inglés. En esta selección se detalla cómo las Expresiones de Quepy se traducen a las partes integrantes de una consulta formal.

A grandes rasgos, Quepy utiliza dos etapas que traducen una pregunta a una Expresión y luego utilizan la Expresión para formar consultas. Esto es así ya que permite soportar diversos lenguajes de consultas. Estas representaciones internas son generales y pueden generar cualquier consulta. Es el programador quien se encarga de especificar las reglas de construcción de las expresiones y las de traducción a lenguaje formal, por ejemplo SPARQL.

2.2.1. Construcción de las consultas

Para entender mejor cómo funciona Quepy internamente veamos en ejemplo en particular, extraído de la documentación oficial ⁴. Este ejemplo corresponde a una aplicación realizada para generar consultas SPARQL para ser enviadas a un motor de la DBPedia. Analicemos primero cómo se definen los elementos del DSL para luego seguir con las plantilla propiamente dichas.

Ejemplo 3. Definición de un elemento del DSL.

```
from quepy.dsl import FixedRelation

class IsDefinedIn(FixedRelation):
    relation = "rdfs:comment"
    reverse = True
```

⁴<http://quepy.readthedocs.org/en/latest/tutorial.html>

La clase *IsDefinedIn* es una Expresión que representa una relación entre dos objetos, como vimos anteriormente en RDF. Dependiendo del lenguaje de consulta tendrá distintas traducciones, y en particular para SPARQL es equivalente a:

```
?target rdfs:comment ?definition
```

donde *?target* y *?definition* son parámetros que tomará la Expresión al instanciarse.

Las expresiones pueden construirse progresivamente a partir de otras expresiones como veremos a continuación.

Ejemplo 4. Plantilla para las preguntas de tipo “What is ... ?”.

```
from refo import Group, Question
from quepy.dsl import HasKeyword
from quepy.parsing import Lemma, Pos, QuestionTemplate

from dsl import IsDefinedIn

class WhatIs(QuestionTemplate):

    aux = Question(Pos("DT")) + Group(Pos("NN"), "target")
    regex = Lemma("what") + Lemma("be") + aux + Question(Pos("."))

    def interpret(self, match):
        thing = match.target.tokens
        target = HasKeyword(thing)
        definition = IsDefinedIn(target)
        return definition
```

Observemos que la clase tiene un atributo llamado *regex* que corresponde a la expresión regular que define la plantilla. Profundizaremos en la estructura de estas expresiones regulares más adelante, pero ahora notemos que uno de los elementos tiene una etiqueta *target*. Si la pregunta ingresada por el usuario concuerda con esta expresión regular, entonces los elementos que concuerden con las sub expresiones etiquetadas serán pasados al método *interpret* de la clase. En este caso, el segmento de oración que corresponda a *Group(Pos("NN"))* (un conjunto de sustantivos) será un atributo del parámetro *match* recibido por *interpret*.

El método *interpret* construye una Expresión de tipo *HasKeyword* a partir de *target* y luego la utiliza para contruir otra Expresión de tipo *IsDefinedIn*. El resultado final de la Expresión traducida a SPARQL para la pregunta “What is a car?” será:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX quepy: <http://www.machinalis.com/quepy#>
```

```
SELECT DISTINCT ?x1 WHERE {
    ?x0 quepy:Keyword "car".
    ?x0 rdfs:comment ?x1.
}
```

2.2.2. Plantillas y sus expresiones regulares

Describiremos a continuación más en detalle la estructura de las plantillas que permiten crear una Expresión a partir de una pregunta. Cada una de las plantillas está construida en base a la librería REfO⁵, que define expresiones regulares entre objetos complejos de Python, no solamente cadenas de caracteres.

Ejemplo 5. Expresión regular del ejemplo 4⁶.

```
regex = Lemma("what") + Lemma("be") + Question(Pos("DT"))
        + Group(Pos("NN"), "target") + Question(Pos("."))
```

Para analizar si una frase concuerda o no con una expresión regular, Quepy transformará la oración con el analizador sintáctico indicado para obtener el lemma y las etiqueta POS de cada una de sus palabras. Luego, utilizará esa información para compararla con la expresión regular. Entonces, nuestro ejemplo concordará con una frase cuya primera palabra tenga lemma “what”, su segunda palabra tenga lemma “be”, su tercera palabra (opcionalmente) tenga etiqueta POS “DT”, etc.

Dada una pregunta, Quepy intentará encontrar una concordancia con cada una de estas expresiones regulares existentes. Si la encuentra, entonces utilizará el método *interpret* que explicamos en la sección anterior para construir una Expresión y luego una consulta.

Definir patrones de reconocimiento de esta manera permite representar tanto información semántica como sintáctica y por lo tanto tienen mayor poder expresivo que cualquiera de ellas por separado. Veremos más adelante que elegimos estos patrones como una forma de representación de las preguntas para clasificar.

⁵<https://github.com/machinalis/refo>

⁶Reemplazamos la variable aux por su contenido para mayor claridad, lo cual no afecta el significado de la expresión regular.

2.3. Formalización del problema

A pesar de las numerosas ventajas de Quepy, también existen desventajas. La más importante de ellas es que, al utilizar expresiones regulares, los patrones no tienen flexibilidad y dependen fuertemente del analizador sintáctico y POS tagger que utilicen.

En particular, si tomamos el ejemplo de la sección anterior, no se podrían reconocer preguntas del estilo “Definition of a car” o “How would you define what a car is?”. La respuesta a estas preguntas se obtiene con la misma consulta generada que acabamos de analizar, por lo cual son esencialmente equivalentes. Diremos entonces que estas preguntas comparten la misma semántica, y que son reformulaciones una de la otra.

Para agregar un nuevo tipo de pregunta al sistema se deben definir sus patrones y sus traducción a una consulta. Gracias a la gran cantidad de formas distintas en las que se puede expresar una pregunta es imposible construir todas las expresiones regulares necesarias, y los sistemas de Quepy están fuertemente limitados por esta característica. Si los patrones fueran más generales o pudieran inferirse de alguna forma, entonces ampliar los tipos soportados consistiría sólo en el segundo paso.

Unger et al. [2014] clasifica los sistemas de respuesta a preguntas sobre datos enlazados (QALD por sus siglas en inglés) según sus estrategias de resolución de la pregunta. Entre ellos se encuentra la clase a la cual pertenece Quepy, llamada por los autores “Template-Based approaches” o Aproximaciones Basadas en Patrones. Claramente, la falta de cobertura sobre el universo posible de preguntas en una dolencia de cualquier sistema que utilice patrones estáticos para clasificar las preguntas en una determinada representación.

Lo que nos proponemos entonces lograr con este trabajo es ampliar la cobertura de un sistema QALD basado en concordancia con patrones para reconocer preguntas semánticamente equivalentes a una de sus clases ya definidas en él. El sistema QALD que tomamos como base es Quepy, en particular una aplicación realizada como demostración del producto⁷. A partir de este punto, utilizaremos la palabra Quepy para referirnos tanto al marco de trabajo como a las aplicaciones construidas por él, y en particular a la que estaremos usando.

⁷Puede utilizarse online ingresando a <http://quepy.machinalis.com/>

2.4. Solución propuesta

Como la generación de nuevas plantillas manualmente no es viable, entonces proponemos una solución automática: agregar al sistema un clasificador que identifique (si existiera) el patrón que corresponde a la pregunta. Es tarea del clasificador asociar reformulaciones que tengan la misma semántica a solo patrón. Una vez obtenida la clase semántica e identificado el objeto de la pregunta, Quepy u otro sistema puede construir la consulta directamente. Dejaremos como trabajo futuro el reconocimiento de la entidad base y nos centraremos en la clasificación de las preguntas.

Este enfoque de encontrar reformulaciones de una misma pregunta está enmarcado dentro del reconocimiento de implicaciones textuales y ha sido utilizado previamente para sistema de respuesta a preguntas del usuario. Ou and Zhu [??] utilizan esta técnica tomando como base preguntas modelo construidas automáticamente desde la ontología, y se centran también en la composición de patrones simples para formar otros más complejos. Sin embargo, se limitan a un dominio muy restringido que permite formar texto en lenguaje natural desde las relaciones formales entre las entidades, lo cual sería dificultoso en ontologías complejas como FreeBase. Wang and Li [??] explican otros posibles usos de identificar estas relaciones entre las preguntas para sugerencia de preguntas relacionadas o útiles para el usuario. El trabajo de Koseim and Yousefi [2008], por otra parte, utiliza la reformulación para obtener patrones semánticamente equivalente, pero utiliza durante el entrenamiento del clasificador la respuesta de la pregunta.

Nuestro trabajo será construir y entrenar un clasificador capaz de recibir una pregunta y decidir a qué clase semántica pertenece, siguiendo la definición de Sebastiani [2002]:

Definición 1. La clasificación de una instancia es la asignación de un valor booleano a cada par $\langle x_i, c_j \rangle \in \mathcal{X} \times \mathcal{C}$, donde \mathcal{X} es el dominio de las instancias y \mathcal{C} es el conjunto de clases posibles.

Asignaremos el valor *Verdadero* a los pares $\langle x_i, c_i \rangle$ si la clase c_j corresponde a la instancia x_i , y *Falso* en el caso contrario.

bla bla bla es necesaria la formalización así? Falta la definición de entrenamiento.

\mathcal{C} para esta clasificación es el conjunto de clases semánticas de Quepy, es decir, cada una de las plantillas o patrones. El ejemplo que describimos en la sección anterior corresponde a la clase “What is”. Todas las preguntas que puedan responderse a través de la consulta generada por esta plantilla serán clasificadas dentro de esta clase. La cantidad total de clases es 29, las agregamos en un APENDICE?

Aunque la tarea a realizar no parece compleja y ha sido ampliamente estudiada, nos encontramos con numerosos obstáculos que impiden utilizar algún método estándar de clasificación de texto. A continuación discutiremos dos de estos inconvenientes y las decisiones que tomamos para resolverlos.

2.4.1. De la Clasificación Supervisada al Aprendizaje Activo

En primer lugar, no contamos con un corpus anotado que permita utilizar clasificación supervisada común. Desarrollamos entonces un pequeño corpus a partir de los ejemplos incluidos en Quepy. Por cada clase agregamos también algunos casos de reformulaciones no reconocidos por la aplicación y también las etiquetamos. El resultado final fueron 106 preguntas, un número más que modesto y que difícilmente cubre el universo de posibles reformulaciones de todas las clases.

Sin embargo, existen numerosos corpus de preguntas utilizados para otras tareas de clasificación que no están etiquetados. Por lo tanto, decidimos utilizar un enfoque semiautomático que comience con un conjunto de semillas y que utilice las preguntas no etiquetadas paulatinamente para aprender la clasificación. Esto nos permitirá compensar la falta de cobertura sobre el dominio.

La fuente más importante de preguntas para la construcción del corpus no anotado fueron las preguntas de entrenamiento y evaluación de las tareas del TREC⁸ desde el año 2008. Por lo tanto, consideramos que nuestro conjunto representativo de las posibles preguntas que un usuario podría esperar que un sistema responda. Sin embargo sólo una porción muy pequeña de ellas se corresponde con alguna de las clases de Quepy. Por lo tanto, entrenar un clasificador con tan alta cantidad de ruido sería una tarea muy difícil.

Tengamos en cuenta también que los límites de una clase semántica no siempre están claros y algunas veces dependen fuertemente de la estructura de los datos en la ontología.

Ejemplo 6.

1. “What is the tallest mountain?”
2. “What is the Everest mountain?”

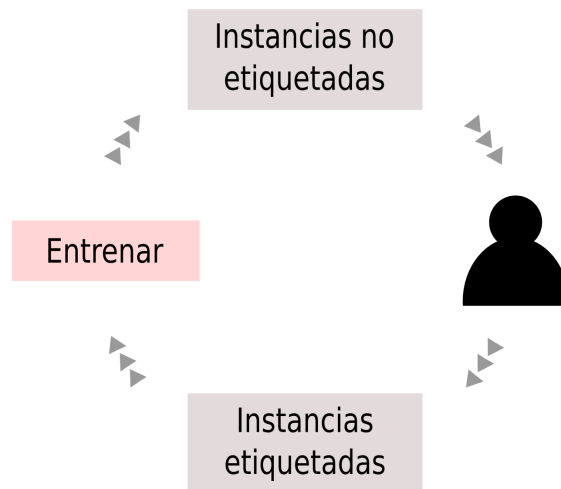
Estas preguntas son muy similares, y sin embargo sólo la segunda pertenece a la clase “whatis”: para responder la primera pregunta debe obtenerse la altura de todas las montañas de la base de datos y seleccionar la mayor.

⁸<http://trec.nist.gov/data/qamain.html>

Por este motivo decidimos utilizar una plataforma de aprendizaje activo donde un oráculo humano ayudará al sistema a delimitar estas sutilezas semánticas. Hospedales et al. [2011] y Ertekin et al. [2007] describe clasificadores adaptados a través del aprendizaje activo para encontrar y clasificar ejemplos raros de clases minoritarias. Además de ello, el aprendizaje activo es una estrategia que obtiene buenos resultados para problemas con una gran cantidad de clases, de acuerdo con Jain and Kapoor [2009].

Settles [2009] explica que el aprendizaje activo es un paradigma donde el aprendiz selecciona preguntas para que un humano u oráculo las etiquete. Si el aprendiz elige las instancias de las cuales puede aprender más información, entonces se minimiza la cantidad de instancias etiquetadas necesarias para lograr el mismo desempeño. La mayor motivación para este tipo de estrategias es la dificultad de conseguir dato etiquetados al mismo tiempo que se disponen de grandes cantidad de ejemplos no etiquetados, tal y como es nuestro caso. Utilizaremos aprendizaje activo para conseguir un corpus etiquetado de entrenamiento con el menor esfuerzo posible.

Figura 2.1: Arquitectura de un sistema de aprendizaje activo



2.4.2. Aprendizaje activo sobre características

Durante las primeras etapas de desarrollo y especificación del problema debimos definir la representación de las instancias ante el clasificador. Sin embargo, al no existir trabajos previos con la misma aproximación al problema no tenemos un punto de referencia para tomar como ejemplo. Por esto, decidimos incluir características

tentativamente y realizar experimentos con aprendizaje activos sobre características e instancias.

En un enfoque como este se pedirá al usuario que etiquete las características seleccionadas con una clase si la presencia de la característica en una instancia es indicativa de la clase. Druck et al. [2009] han realizado experimentos sobre clasificación de texto en donde se demuestra que el aprendizaje activo sobre características puede dar mejores resultados incluso que el aprendizaje sobre instancias. Durante este trabajo ayudará a identificar las características que mejor describen las instancias para la clasificación descripta.

Las etiquetas obtenidas se utilizarán también para entrenar el clasificador reforzando la probabilidad de una característica dada una clase, como veremos en detalle en la implementación.

2.4.3. Dualist

Dualist es un sistema muy similar al que estamos planteando desarrollado por Settles [2011] que combina el aprendizaje sobre instancias y sobre características para diversas tareas como el análisis de sentimientos o la clasificación de documentos.

Nuestra implementación seguirá los mismos lineamientos principales y pondremos en práctica las técnicas de selección de ejemplos con las de entrenamiento del clasificador para obtener resultados sobre nuestra configuración. Decidimos tomarlo como modelo ya que se centra en la interfaz con el usuario y en la capacidad del software de ser utilizado en tiempo real. Nosotros deseamos lograr un sistema que sea viable de integrar con Quepy y complementar aplicaciones reales, en lugar de ser utilizado sólo para demostrar los resultados de este trabajo.

En esencia, Dualist es un marco de trabajo para construir aplicaciones de aprendizaje activo donde los usuarios pueden etiquetar ejemplos o características.

Figura 2.2: Captura de pantalla de la interfáz gráfica de Dualist.



Capítulo 3

Representación de las preguntas

Una importante parte de cualquier problema que aborde el lenguaje natural es la representación del mismo. Las características relevantes son propias de cada problema, si bien existen numerosos ejemplos bibliográficos a tomar como modelo. El siguiente paso es identificar qué características de la pregunta son indicativas de su clase semántica, lo cual no tiene una respuesta intuitiva.

Nuestra primera aproximación fue utilizar criterios estándares en la categorización de texto como los lemmas de las palabras y sus etiquetas POS. Sebastiani [2002] describe que a pesar de su simpleza son representaciones muy poderosas y que otras más complejas no necesariamente llevarán a un mejor desempeño del clasificador.

3.1. Subpatrones

Como presentamos en el ejemplo 6 algunas preguntas tienen tanto lemmas como etiquetas POS muy similares y sin embargo pertenecen a clases distintas. La forma en la que Quepy distingue estos casos es utilizando la estructura de la frase, representada en sus patrones. Por eso, decidimos utilizar además como característica los patrones que concuerdan con la pregunta.

Esta representación por sí sola tampoco mejora nuestra situación inicial, ya que sólo reconoce las preguntas que corresponden a un patrón. La solución que encontramos para este problema fue dividir cada patrón en todos sus posibles subpatrones y asignar a cada instancia la concordancia con los subpatrones.

Ejemplo 7. Subpatrones del ejemplo 5.

1. `Lemma("what") + Lemma("be") + Question(Pos("DT"))`
+ `Group(Pos("NN"), "target") + Question(Pos("."))`

2. $\text{Lemma}(\text{"what"}) + \text{Lemma}(\text{"be"}) + \text{Question}(\text{Pos}(\text{"DT"}))$
 $+ \text{Group}(\text{Pos}(\text{"NN"}), \text{"target"})$
3. $\text{Lemma}(\text{"what"}) + \text{Lemma}(\text{"be"}) + \text{Question}(\text{Pos}(\text{"DT"}))$
4. $\text{Lemma}(\text{"what"}) + \text{Lemma}(\text{"be"})$
5. $\text{Lemma}(\text{"what"})$
6. $\text{Lemma}(\text{"be"})$
7. $\text{Lemma}(\text{"be"}) + \text{Question}(\text{Pos}(\text{"DT"}))$
 $+ \text{Group}(\text{Pos}(\text{"NN"}), \text{"target"}) + \text{Question}(\text{Pos}(\text{"."}))$
8. ...

3.2. Nombres y tipos de entidades

Existe un tipo más de característica que determina fuertemente la semántica de la pregunta. La forma de acceder a una propiedad de una entidad dentro de una base de datos está íntimamente ligada a la estructura que le dió el usuario al ingresar los datos. Por lo tanto, la misma propiedad como “fecha de creación” puede tener dos nombres distintos para tipos de entidades. Por lo tanto, dependiendo de un tipo o de otro la consulta que debe generarse es distinta, y por eso son necesarias dos clases semánticas distintas.

Ejemplo 8. Ejemplos con semántica diferenciada por el tipo de la entidad.

1. “Who are the actors of Titanic?”
2. “Who are the actors of Friends?”

Para obtener los actores que trabajaron en una película, como en el caso de la primera pregunta, debe utilizarse la relación “/film/film/starring”, mientras que en el caso de una serie televisiva se utiliza “/tv/tv_program/regular_cast”.

El único indicio de la clase semántica en estos casos es la entidad referenciada. Por ello, la incluimos como una característica más. Adicionalmente, agregaremos los tipos de dicha entidad en la base de conocimiento, particularmente para esta aplicación FreeBase. Cabe destacar que no todas las preguntas tienen una entidad, y en el caso

de que sí tenga no siempre podemos reconocerla. Esto depende del sistema externo de reconocimiento de nombres de entidades o NER por sus siglas en inglés.

En resumen, las características propuestas para el sistema son:

- Etiquetas POS.
- Lemmas.
- Concordancias a patrones parciales.
- Nombre de la entidad referenciada.
- Tipos de la entidad referenciada.
- Bigramas y trigramas.
- Bigramas mezclando Lemmas y POS.

Capítulo 4

Arquitectura del sistema

Si bien el objetivo principal de este trabajo es incrementar la cobertura de Quepy, deseamos también que el sistema esté compartimentado de tal forma que sus componentes puedan utilizarse individualmente para otras aplicaciones.

Trabajamos sobre dos grandes áreas: un clasificador automático entrenado utilizando aprendizaje activo sobre instancias y características, y la representación de las preguntas para maximizar los resultados del clasificador. El clasificador puede ser visto como una tarea general, y por lo tanto lo desarrollamos como una librería portable cuyos parámetros pueden ser definidos por el usuario. Sin embargo, la representación de las preguntas ya descrita es algo completamente ligado al sistema de preguntas y por ello lo diseñamos como una extensión opcional de Quepy que puede utilizar cualquier clasificador.

A continuación explicaremos el ciclo de aprendizaje activo y cada una de estos módulos.

4.1. Framework para aprendizaje automático sobre instancias y características

4.1.1. Funcionalidad

4.1.2. Parámetros

Clasificador El usuario debe definir qué clasificador de sklearn utilizar.

Características

Corpus

Función de representación al usuario El usuario debe proveer una interfáz gráfica para presentar los datos al usuario y obtener una respuesta.

Si bien ésta aproximación es altamente paramétrica, cabe destacar que permite gran flexibilidad con respecto a los datos ingresados. Al permitir elegir tanto características como el corpus y el clasificador, puede ser utilizado dentro de cualquier ámbito incluso no relacionado al procesamiento del lenguaje natural.

4.1.3. Selección de instancias

4.1.4. Selección de características

En Dualist se presentaban al usuario dos listas de posibles características ordenadas de acuerdo a la correlación que tenían con cada clase. La principal diferencia con nuestra arquitectura es que tenemos 30 clases, no sólo dos, y por lo tanto debemos cambiar la forma de interacción. Por lo tanto no sólo debemos utilizar conceptos del aprendizaje activo para la selección de instancias y características, sino sobre las clases que vamos a mostrar. Cabe destacar también que en Dualist el active learning sobre ejemplos es totalmente independiente del active learning sobre instancias.

Formas para ordenar las clases para preguntar al oráculo.

- Clases que tengan la mayor probabilidad para un feature cualquiera.
- Clases que tengan la mayor suma de probabilidades sobre todos sus features.
- La clase con mayor probabilidad.
- La clase con menor probabilidad.
- La clase con mayor cantidad de instancias anotadas.
- La clase con menor cantidad de instancias anotadas.

4.1.4.1. Information Gain

Un criterio ampliamente utilizado para la selección de features es Information Gain.

Definición wikipedia: In probability theory and information theory, the Kullback–Leibler divergence[1][2][3] (also information divergence, information gain, relative entropy, or KLIC; here abbreviated as KL divergence) is a non-symmetric measure of the difference between two probability distributions P and Q .

Danny Roobaert and Chawla [??] Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy

4.1.5. Maximización de la esperanza

El algoritmo de maximización de la esperanza es ampliamente utilizado para inferir parámetros desconocidos en una distribución que tiene estados latentes. Para esto, utiliza estimadores de máxima verosimilitud, y para clasificación el estado latente al que nos referimos es la misma clase. A grandes rasgos, el proceso funciona en dos pasos E y M, por sus siglas en inglés Expectation y Maximization. El paso E es el que calcula el valor esperado de la verosimilitud asumiendo que la distribución actual es verdadera y no existen variables no observadas. Para realizarlo, utilizamos el clasificador en su estado actual para etiquetar probabilísticamente el conjunto de datos no etiquetados, y asumimos dichas etiquetas como correctas. Con este conjunto de nuevos datos se calcula la verosimilitud del modelo. Recordemos que la verosimilitud es una función sobre los parámetros del modelo y un conjunto de datos que calcula la probabilidad de los valores tomados por cada variable aleatoria del modelo bajo dichos parámetros. Si los parámetros actuales fueran correctos, entonces la verosimilitud obtenida sería la verosimilitud real del modelo, pero si no lo son provee una cota inferior. Luego, como tenemos una función sin valores no observados, el paso M maximiza la función encontrando parámetros del modelo más adecuados. De esta forma optimizamos la cota inferior encontrada. Este paso se repite numerosas veces hasta lograr convergencia. Sin embargo tomamos ejemplo de Settles y realizamos una sola iteración, dado que argumenta que las siguientes iteraciones no aportan significativamente a la precisión del clasificador.

Capítulo 5

Entorno de experimentación

5.1. Ejemplos seleccionados

Distribución actual del corpus: Quepy questions 115 Recognized 58 Unrecognized 57 Other questions 6658 Labeled 607 Unlabeled 6051

Test corpus 250 Training corpus 165 Unlabeled corpus 6358

5.1.1. Proceso previo

Al momento de extraer las características que mencionamos en la sección anterior encontramos varios problemas. Una aproximación simple al aprendizaje activo incluye reentrenar el clasificador en cada una de las iteraciones del ciclo, cambiando así el modelo. Al introducir el etiquetado de características ya no se puede cambiar el modelo sin perder rastro de la ubicación de las características etiquetadas dentro de las matrices internas del clasificador. Por esto es que tuvimos que cambiar la implementación básica y extraer todas las características dentro de la instancia de preproceso. De esta forma, nuestras matrices iniciales tienen toda la información tanto del corpus anotado como no anotado.

Si bien las características anteriores pueden tomar valores en un amplio rango de enteros, decidimos sólo medir la presencia o ausencia de cada una de ellas. Por un lado, consideramos que el usuario no debería etiquetar cada característica en base a su valor, ya que este no es distintivo de la clase. Por ejemplo, si el lemma de una palabra como “movie” aparece más de 3 veces o menos no influye en las etiquetas que puedan asignarse a ellas.

5.2. Usuarios Emulados

Para poder realizar experimentos automáticos sin que el usuario tenga que ingresar la misma información repetidas veces decidimos guardar las respuestas obtenidas. Por ello agregamos etiquetas al corpus no etiquetados, que por supuesto no son consultadas, y creamos un corpus de asociaciones características-clases. El corpus de características no es más que una matriz ternaria con tres valores posibles: asociación positiva, no asociación, desconocido. De esta forma también evitamos preguntar a un usuario por las características que ya han sido vistas pero no están relacionadas a ninguna clase en particular. Observemos que esta forma de guardar la información soporta etiquetas múltiples.

5.3. Experimentos realizados

5.3.1. Métricas utilizadas

Exactitud Llamamos exactitud a la cantidad de preguntas etiquetadas correctamente sobre el total de preguntas clasificadas.

Curva de aprendizaje Definimos la curva de aprendizaje como la exactitud del clasificador en función de la cantidad de ejemplos o características etiquetados necesarios.

Coeficiente Kappa de Cohen Esta medida ajusta la exactitud del clasificador obtenido a la de un clasificador aleatorio. Una exactitud del 80 % no es muy sorprendente si asignando etiquetas al azar obtenemos una exactitud del 70 %. En nuestro caso el corpus de evaluación contiene aproximadamente un 75 % de instancias de clase “otro”, por lo tanto un clasificador que elija esta etiqueta todas las veces obtendría una exactitud semejante. Esta métrica nos permitirá superar este inconveniente y obtener valores más reales.

Una definición más formal del Coeficiente de Kappa es la que propone Carletta [1996]:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

donde $P(A)$ es la proporción de veces que los clasificadores acuerdan y $P(E)$ es la proporción de veces que se esperaría que acuerden por casualidad. En este caso, uno de los clasificadores es el Multinomial Bayesiano entrenado y el otro son las etiquetas del corpus de evaluación. Por lo tanto, $P(A)$ no es otra cosa

más que la exactitud calculada en el primer item. Adicionalmente, calculamos $P(E)$ de la siguiente forma:

$$P(E) = \frac{\sum_{i \in \mathcal{C}} Pr(\hat{x}_i) * Pr(x_i)}{|\mathcal{E}|}$$

donde \mathcal{C} es el conjunto de clases, \mathcal{E} es el corpus de evaluación, $Pr(\hat{x}_i)$ es la proporción de instancias etiquetadas por el clasificador con la clase i , y $Pr(x_i)$ es la proporción de instancias que pertenecen realmente a la clase i .

Precisión y exhaustividad por clase Estas dos medidas pueden utilizarse sólo en clasificación binaria, por lo que tomaremos sus valores para cada una de las clases posibles. Definimos precisión como la cantidad de instancias etiquetadas para una clase que son correctas (positivos verdaderos o P_v) sobre la cantidad de instancias etiquetadas para esa clase (P_v y falsos positivos o P_f).

$$Precision(C_i) = \frac{P_v}{P_v + P_f}$$

La exhaustividad, por otro lado, está definida como la cantidad de instancias etiquetadas correctamente (P_v) de una clase dada sobre la cantidad de instancias que pertenecen a la clase verdaderamente (P_v y falsos negativos o N_f).

$$Exhaustividad(C_i) = \frac{P_v}{P_v + N_f}$$

5.3.2. Baseline

Tomaremos como baseline dos métodos:

Selección de instancias y características al azar

Los experimentos que solicitan al oráculo información sobre instancias y características se realizaron alternando pocas preguntas relativas a instancias y la misma cantidad relativas a características. Si bien los dos ciclos de etiquetado son completamente independientes en el sistema, tomamos esta decisión para eliminar las diferencias entre experimentos que puedan introducirse a partir de la elección de los usuarios. Otro punto en el que difieren nuestros experimentos de una sesión no simulada con un usuario es la cantidad de veces que se reentrena. Con objeto de obtener una medición precisa de la curva de aprendizaje reentrenamos el clasificador luego de cada ronda instancias-características descriptas anteriormente. Esto es costoso para grandes volúmenes de datos y podría hacer perder al sistema su interactividad.

5.3.3. Hipótesis 1

El aprendizaje activo obtiene mejores resultados que un clasificador normal utilizando la misma cantidad de datos.

Para comprobar esta hipótesis realizamos sesiones de etiquetamiento sobre instancias y características, pero eligiendo al azar cuáles presentar al usuario. Luego compararemos las curvas de aprendizaje de ambos experimentos.

5.3.4. Hipótesis 2

El aprendizaje activo sobre instancias y características obtiene mejores resultados que el aprendizaje activo sobre instancias o características por separado.

5.3.5. Experimento 3

Hipótesis El aprendizaje supervisado sobre instancias y características obtiene mejores resultados que el aprendizaje supervisado sobre instancias, aún si las características son pocas.

5.3.6. Experimento 4

Hipótesis ??.

La idea es ver qué método de selección de features es mejor:

Puede ser trabajo futuro

Resultados

	Clase con probabilidad alta	Clase con probabilidad baja
Característica con probabilidad alta	medicion	medicion
Característica con probabilidad baja	medicion	medicion

	Clase con probabilidad alta	Clase con probabilidad baja
Característica con IG alta	medicion	medicion
Característica con IG baja	medicion	medicion

5.3.7. Experimento 5

Hipótesis Seleccionar features para etiquetar que tengan alta confiabilidad/correlación, y luego de superado un cierto límite pasar a los que tiene baja confiabilidad/correlación permite al clasificador eliminar el ruido no introducido por la baja cantidad de ejemplos y al mismo tiempo expandir la cobertura. Dejar para mas adelante

Experimento Entrenamiento supervisado con y sin etiquetado de features Validacion del etiquetado de features

Experimento 6 Information gain sobre todo el corpus o solo el etiquetado. IG sobre el corpus anotado + frecuencia en no anotado vs IG sobre todo el corpus anotado y no anotado.

Experimento 7 Coocurrencia de features con otros features. (Información mutua) Un feature se rankea mas alto si coocurre con features que se rankean alto. Tomando como base la frecuencia.

Experimento 8 Information gain sirve o alcanza sólo con usar coocurrencia?

Bibliografía

- Tim Berners-Lee. Linked data - design issues, November 2014. URL <http://www.w3.org/DesignIssues/LinkedData.html>.
- Dan Brickley and Ramanathan V. Guha. Rdf vocabulary description language 1.0: Rdf schema - w3c recommendation, November 2014. URL <http://www.w3.org/TR/rdf-schema/>.
- Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.
- Tom Heath Christian Bizer and Tim Berners-Lee. Liked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.
- Grigoris Karakoulas Danny Roobaert and Nitesh V. Chawla. Information gain, correlation and support vector machines. ??, page ??, ??
- Gregory Druck, Burr Settles, and Andrew McCallum. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 81–90, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL <http://dl.acm.org/citation.cfm?id=1699510.1699522>.
- Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*, pages 127–136. ACM, 2007. ISBN 978-1-59593-803-9.
- Poonam Gupta and Vishal Gupta. A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4):1–8, 2012.
- TimothyM. Hospedales, Shaogang Gong, and Tao Xiang. Finding rare classes: Adapting generative and discriminative models in active learning. In JoshuaZhexue

- Huang, Longbing Cao, and Jaideep Srivastava, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6635 of *Lecture Notes in Computer Science*, pages 296–308. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-20846-1.
- P. Jain and A. Kapoor. Active learning for large multi-class problems. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 762–769, June 2009.
- Leila Kosseim and Jamileh Yousefi. Improving the performance of question answering with semantically equivalent answer patterns. *Data Knowl. Eng.*, 66(1):53–67, 2008. ISSN 0169-023X. URL <http://dx.doi.org/10.1016/j.datak.2007.07.010>.
- Shiyan Ou and Zhenyuan Zhu. An entailment-based question answering system over semantic web data. *??, ??(?) : ?–?, ??*
- Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002. ISSN 0360-0300.
- Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.
- Burr Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. *2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, 2011.
- Christina Unger, André Freitas, and Philipp Cimiano. An introduction to question answering over linked data. In Manolis Koubarakis, Giorgos Stamou, Giorgos Stoilos, Ian Horrocks, Phokion Kolaitis, Georg Lausen, and Gerhard Weikum, editors, *Reasoning Web. Reasoning on the Web in the Big Data Era*, volume 8714 of *Lecture Notes in Computer Science*, pages 100–140. Springer International Publishing, 2014. ISBN 978-3-319-10586-4.
- Rui Wang and Shuguang Li. Constructing a question corpus for textual semantic relations. *??, ??(?) : ?–?, ??*