Uma abordagem de classificação automática para Tipo de Pergunta e Tipo de Resposta

Patricia Nunes Gonçalves[1], **António Horta Branco**[1]

[1]Faculdade de Ciências da Universidade de Lisboa
Lisboa - Portugal

{patricia.nunes,antonio.branco}@di.fc.ul.pt

Abstract. *The question type classification and answer type classification are very important tasks for Question Answer Systems. This paper presents an automatic approach using machine learning for these tasks. We used decision trees as machine learning algorithm and 14 features developed using a tagger and a named entity systems.*

Resumo. *A classificação de tipos de pergunta e tipo de resposta são tarefas importantes para Sistemas de Respostas a Perguntas. Este artigo apresenta uma abordagem automática utilizando aprendizado de máquina para estas tarefas. O algoritmo utilizado foi árvores de decisão e foram usados 14 atributos criados através das informações fornecidas por um tagger e um sistema de reconhecimento de entidades mencionadas.*

1. Introdução

Com avanço da internet é cada vez mais comum as pessoas necessitarem buscar informação na Web. A forma mais comum de se fazer pesquisa é utilizando um motor de busca, como por exemplo, o Google[1]. O utilizador quando necessita de alguma informação usa palavras-chave para realizar a busca na internet. Centenas e às vezes milhares de páginas são retornadas e o utilizador realiza a leitura uma-a-uma destas páginas em busca da informação necessária.

Observa-se a tendência das perguntas estarem tornando-se cada vez mais específicas, por exemplo, "Qual o nome da operação para controlar o embargo a armas no conflito da ex-Jugoslávia iniciado em Junho de 1993?". Um simples motor de busca não consegue responder satisfatoriamente a perguntas escritas de forma natural. Sendo assim, há vários anos tem-se estudado sistemas específicos para atender a essa necessidade do utilizador. Estes sistemas são chamados de Sistemas de Respostas a Perguntas - SRP. Num Sistema de Respostas a Perguntas as questões são feitas em linguagem natural e o sistema busca automaticamente a resposta [Paşca 2003].

A arquitectura básica de um SRP baseia-se em três grandes módulos: *Processamento da pergunta*, módulo responsável por transformar a pergunta em uma expressão lógica que será usada para buscar os textos relevantes para a pergunta. *Recuperação da informação*, módulo responsável por identificar quais as passagens dos textos selecionados são relevantes para a pergunta e, por último, *Extração da resposta*, módulo responsável por escolher a resposta mais relevante para a pergunta.

[1]http://www.google.com/

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 189–193,
Cuiabá, MT, Brazil, October 24–26, 2011.

Este artigo tem como foco as duas tarefas principais no primeiro módulo de um SRP, nomeadamente, classificação do tipo de pergunta e tipo de resposta. Na Seção 2 descrevemos informações sobre o corpus que foi utilizado. Na Seção 3 discutimos os experimentos realizados e os resultados obtidos. As conclusões são apresentadas na Seção 4.

2. Corpus

O corpus utilizado neste trabalho é composto por 1540 frases interrogativas anotadas manualmente com o tipo de pergunta e tipo de resposta e na Tabela 1 apresenta a estatística do corpus.

Tabela 1. Estatística do corpus para Tipos de Perguntas e Tipos de Respostas

Perguntas		
Tipo	Qtde	Exemplo
O_que	153	O que é a Yves Saint-Laurent?
Que_N	413	Que dinastia governa a Jordânia?
Quem	402	Quem é Nelson Mandela?
Onde	120	Onde morreu Yitzhak Rabin?
Qual	120	Qual é a capital do Chipre?
Quanto	152	Quantos cidadãos europeus há?
Quando	116	Quando nasceu Luc Jouret?
Como	41	Como se pode poupar energia?
Verbal	23	Diga um escritor norueguês.
Total	1540	

Respostas		
Tipo	Qtde	Exemplo
Organização	331	De que jornal é repórter Clark Kent?
Pessoa	466	Quem é o fundador do Greenpeace?
Localização	266	Onde se encontra a Potsdamer Platz?
Número	149	Quantas pessoas vivem em França?
Medida	70	A que velocidade viaja a luz?
Temporal	184	Quando nasceu Christopher Reeve?
Lista	21	Que genes causam o cancro?
Outras	53	Que nome tem o iate real britânico?
Total	1540	

Os tipos de perguntas estão divididos em 9 classes distintas: `O_que`, `Que_N`, `Quem`, `Onde`, `Qual`, `Quanto`, `Quando`, `Como` e `Verbal`. Nota-se que o pronome interrogativo "QUE" difere dos demais e é subdividido em 2 classes `O_que` e `Que_N`. A diferença encontra-se na composição da frase interrogativa, sendo composta pelo artigo "O" seguido do pronome "QUE" (marcação `O_que`). Ou então a frase inicia directamente com o pronome "QUE" sem a presença do artigo e seguido de um nome (marcação `Que_N`). Cabe ainda ressaltar outro tipo de pergunta encontrada na anotação no corpus com tipo `Verbal`. Este tipo é marcado quando não há pronome interrogativo na frase, geralmente este tipo de pergunta inicia com um verbo.

Os tipos de respostas também estão anotados manualmente no corpus e são divididos em 8 classes: `Organização`, `Pessoa`, `Localização`, `Número`, `Medida`, `Temporal`, `Lista` e `Outras`. É importante destacar a diferença entre as classes: `Número` e `Medida`. Quando é realizada a anotação `Número` como tipo de resposta espera-se qualquer número como resposta. No caso da classe `Medida` a resposta esperada é mais complexa que simplesmente um número, mas sim uma medida. Alguns exemplos de respostas mensuráveis são: "Qual a altura...?, Quantos quilos...?, Qual o tamanho...?, Qual a velocidade...?, Qual a distância...?", entre outras.

3. Experimentos

Para a realização dos experimentos, primeiramente, o corpus foi processado usando as ferramentas: LX-Suite [Branco e Silva 2006] e LX-Ner [Ferreira et al. 2007]. O LX-Suite é um *tagger* para o Português que fornece informações no nível da palavra como: etiqueta *part-of-speech*, forma canónica (lema), gênero e número e flexão verbal. O LX-Ner é um sistema de reconhecimento de entidades mencionadas que realiza marcações em entidades dos tipos: pessoa (PER), organização (ORG), evento (EVT), localização (LOC), trabalho/obra (WRK) e outros (MSC). Todas as ferramentas estão disponíveis online[2].

[2]http://lxcenter.di.fc.ul.pt/

Usando as anotações fornecidas pelas ferramentas descritas acima foram desenvolvidos 14 atributos que foram utilizados para desenvolver os classificadores. Os atributos estão descritos na tabela 2. Os atributos implementados foram extraídos para todas as frases do corpus e usados no treinamento e teste dos classificadores.

Tabela 2. Atributos implementados

Nome	Tipo	Descrição
tem_pron_interr	booleano	Este atributo indica se é encontrado algum pronome interrogativo na frase.
inicia_pron_interr	booleano	Este atributo indica se o pronome interrogativo é encontrado no início da frase
ha_verbos	booleano	Este atributo indica quando são encontrado verbos na frase.
verbo_singular	booleano	Este atributo indica se o verbo encontrado está no singular.
inicia_verbo	booleano	Este atributo indica se a frase inicia com verbo.
ha_ent_mencionada	booleano	Este atributo indica se é encontrada alguma entidade mencionada na frase.
numeral	booleano	Este atributo indica se há algum numeral na frase.
adjetivo	booleano	Este atributo indica se há algum adjetivo na frase.
verbo_ser	booleano	Este atributo indica se há o verbo "SER" na frase.
qtde_palavras	numérico	Número de palavras na frase
qtde_verbos	numérico	Número de verbos na frase
qtde_nomes	numérico	Número de nomes na frase - nomes próprios e comuns
pron_interrog	string	Este atributo indica o pronome interrogativo encontrado na frase.
tp_ent_mencionada	string	Este atributo indica qual tipo de entidade mencionada na frase.

3.1. Resultados

Os experimentos foram realizados utilizando a ferramenta Weka [Witten e Frank 2005] que implementa vários algoritmos de aprendizagem automática. O algoritmo escolhido foi árvores de decisão com *10-fold-cross-validation*. Os resultados obtidos com os classificadores para cada uma das tarefas são apresentados a seguir.

A tarefa de classificação de tipo de perguntas constitui na classificação automática das perguntas em 9 classes (apresentadas na Seção 2) e os resultados são apresentados na Tabela 3. Para a *baseline* utilizou-se apenas o pronome interrrogativo como atributo e observamos que o classificador não foi suficiente para classificar as classes `O_que` e `Verbal`. A classe `O_que` acabou por ser classificada em `Que_N` por ter o mesmo pronome interrogativo. Já a classe verbal foi erroneamente distribuída nas classes `Onde`, `Qual` e `Quanto`. Usando o classificador com os 14 atributos (descritos na Tabela 2) já foi possível distinguir todas as 9 classes com alto desempenho. Os atributos mais relevantes para esta tarefa são: `ha_verbos`, `numeral`, `adjetivo`, `verbo_ser`, `qtde_nomes` e `pron_interrog`.

Tabela 3. Classificação Tipos de Perguntas

	baseline			14 atributos		
Tipo de Pergunta	Precisão	Cobertura	F-Measure	Precisão	Cobertura	F-Measure
O_Que	0	0	0	0.99	0.85	0.91
Que_N	0.73	0.98	0.83	0.93	0.98	0.96
Quem	1	1	1	1	1	1
Onde	0.98	1	0.99	0.98	1	0.99
Qual	0.98	1	0.99	1	1	0.99
Quanto	0.85	1	0.92	0.91	0.94	0.93
Quando	1	1	1	1	1	1
Como	1	1	1	1	1	1
Verbal	0	0	0	0.8	0.7	0.74
Total	0.79	0.88	0.83	0.97	0.97	0.97

Para a *baseline* da tarefa de classificação de tipo de respostas usamos apenas o pronome interrogativo como atributo. Os resultados dos classificadores são apresentados na Tabela 4. Apesar que no geral a *baseline* e o experimento com 14 atributos terem alcançado valores semelhantes, a *baseline*, somente com o pronome interrogativo, não conseguiu classificar nenhum elemento para as classes `Medida`, `Lista` e `Outras`.

O classificador com todos atributos implementados neste trabalho obteve melhores resultados além de conseguir diferenciar os elementos em todas as classes apresentadas. As classes em que se obteve maior dificuldade na classificação foram: `Organização`, `Medida`, `Lista` e `Outras`. A classe `Organização` teve os seus elementos distribuídos de forma irregular por todas as demais classes. Para a classe `Medida` aconteceu o mesmo, entretanto um maior número de elementos foi classificado erroneamente nas classes `Localização` e `Organização`. As classes `Lista` e `Outras` são as mais difíceis de classificar pois a classe `Lista` pode ser uma lista de pessoas, lista de organização e assim por diante. Já a classe `Outras` é um tipo de tal forma tão genérico que torna difícil distingui-la das outras classes. Os dados para ambas as classes estão esparsos na classificação, entretanto observou-se uma tendência de classificá-las erroneamente na classe de `Organização`. Os atributos mais relevantes para esta tarefa são: `qtde_nomes`, `pron_interrog` e `tp_ent_mencionada`.

Tabela 4. Classificação de Tipos de Respostas

	baseline				14 atributos		
Tipo de Resposta	Precisão	Cobertura	F-Measure		Precisão	Cobertura	F-Measure
Organização	0.51	0.86	0.64		0.58	0.69	0.63
Pessoa	0.94	0.89	0.91		0.91	0.9	0.91
Localização	0.64	0.61	0.62		0.67	0.71	0.69
Número	0.76	0.87	0.81		0.82	0.82	0.82
Medida	0	0	0		0.5	0.4	0.44
Temporal	1	0.63	0.77		0.83	0.71	0.77
Lista	0	0	0		0.29	0.19	0.23
Outras	0	0	0		0.19	0.09	0.13
Total	0.70	0.72	0.69		0.73	0.73	0.73

4. Conclusão

Este trabalho apresentou dois classificadores para as tarefas de classificação automática de tipo de pergunta e tipo de resposta. O objetivo deste trabalho foi a criação de classificadores genéricos capazes de lidar com a diversidade e generalidade da língua Portuguesa e explorar uma abordagem diferente daquela frequentemente usada no estado da arte, que utiliza padrões por regras para encontrar a classificação correcta. Para os classificadores desenvolvidos escolheu-se o algoritmo de árvores de decisão como técnica de aprendizado automática. Entretanto, outros algoritmos de aprendizado automático foram testados e os resultados foram inferiores ou iguais aos alcançados com árvores de decisão.

Os resultados obtidos neste trabalho mostram que métodos superficiais como os que foram desenvolvidos alcançam bons resultados para certos tipos de classes. No entanto, para as classes mais complexas será necessário explorar o uso de atributos linguisticamente ricos. A intenção é estender os classificadores desenvolvidos utilizando outras ferramentas linguísticas, por exemplo, um parser sintático e um parser de dependências através dos quais poderão ser obtidos atributos de maior complexidade linguistica. Outro ponto de expansão deste trabalho é aumentar o corpus com mais frases interrogativas e ainda melhorá-lo com a inclusão de outros tipos de frases não encontradas na atual versão do corpus. Para além disso, a anotação pode ser refinada. A classe `Outras` poderá ser especificada e subdividida e a classe `Lista` será melhorada acrescentando o tipo de lista que se pretende encontrar como resposta. Os resultados obtidos dos classificadores implementados neste trabalho serão parte integrante da nova versão do Sistema XisQuê [Branco et al. 2008] que vem sendo desenvolvida no grupo NLX[3].

[3]http://nlx.di.fc.ul.pt/

Referências

António Branco, Lino Rodrigues, João Silva e Sara Silveira (2008). Xisquê: An online QA Service for Portuguese. Em *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR'08, páginas 232–235, Berlin, Heidelberg. Springer-Verlag.

António Branco e João Ricardo Silva (2006). A suite of shallow processing tools for portuguese: Lx-suite. Em *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters; Demonstrations*, EACL'06, páginas 179–182, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eduardo Ferreira, João Balsa e António Branco (2007). Combining rule-based and statistical methods for named entity recognition in portuguese. Em *Proceedings of the 5th Workshop em Tecnologia da Informação e da Linguagem Humana*, TIL'07, páginas 1615–1624, Porto Alegre, Brasil. Sociedade Brasileira de Computação.

Marius Paşca (2003). *Open-Domain Question Answering from Large Text Collections*. CSLI Studies in Computational Linguistics. CSLI, Stanford, California.

Ian Witten e Eibe Frank (2005). *Data mining: Pratical Machine Learning Tools and Techniques*. Morgan Kaufman Publisher.