Extração de Relações em Títulos de Notícias Desportivas

António Paulo Santos[1], Carlos Ramos[1], Nuno C. Marques[2]

[1] GECAD, Instituto Superior de Engenharia do Porto, Portugal
[2] DI-FCT, Universidade Nova de Lisboa, Monte da Caparica, Portugal
pgsa@isep.ipp.pt, csr@isep.ipp.pt, nmm@di.fct.unl.pt

Resumo Este artigo apresenta e avalia um sistema para extrair relações a partir de títulos de notícias. O sistema não requer que seja definido antecipadamente o conjunto de relações a extrair. Para tal, extrai relações do tipo *(sujeito, verbo, objeto)*. São também extraídos determinados *atributos* dos elementos da relação, assim como a *inter-relação* entre duas relações adjacentes. Os desafios e limitações do sistema são discutidos.

Keywords: extração de relações

1 Introdução

Neste trabalho é apresentado um estudo empírico sobre extração de relações a partir títulos de notícias de desporto em Português. Está-se interessado em extrair relações do tipo *(sujeito, verbo, objeto)* e para cada um dos seus elementos, um *conjunto de atributos*. Pretende-se ainda, extrair a *inter-relação* entre cada duas relações adjacentes. Por exemplo, para os títulos de notícias:

1. Chelsea não irá vencer a Premier League
2. Manchester United perde mas continua na liderança

De (1), pretende-se extrair a relação *(Chelsea, vencer [negação=sim], Premier League)*, em que o atributo *"negação"* é importante para capturar o contexto em que *"vencer"* surge. A palavra *"vencer"* que fora do contexto tem uma orientação semântica positiva, nesta frase, aparece negada, tornando-a negativa.

De (2), pretende-se extrair as relações R1:*(Manchester United, perde, null)* e R2:*(Manchester United, continua, liderança)*, assim como a *inter-relação* R3:*(R1, mas, R2)*, capturando assim a interdependência entre R1 e R2.

As contribuições deste trabalho, incluem, a apresentação do sistema *News2Relations* que:

- Extrai relações a partir de títulos de notícias, sem ser necessário especificar antecipadamente o conjunto de relações a extrair.
- Extrai certos atributos (ex. *adjetivos*, *advérbios* e *negação local*) dos elementos principais da relação.
- Extrai a inter-relação entre duas relações adjacentes.

O trabalho aqui descrito enquadra-se num projeto, em que se pretende extrair conhecimento diverso de títulos de notícias desportivas. Esta é a razão que motiva a extração das relações apresentadas neste artigo.

O trabalho está organizado da seguinte forma. Na seção 2 define-se o problema. Na seção 3 descreve-se o trabalho relacionado. Na seção 4 é apresentado a arquitetura de um sistema para extração de relações, que são avaliadas na seção 5.3. Por fim, na seção 6 conclui-se com um breve resumo e trabalho futuro.

2 Definição do Problema

O problema estudado neste trabalho é o da extração de relações do tipo (*sujeito, verbo, objeto*), em que o único elemento obrigatório é o verbo. Cada um destes elementos, se não nulo, pode ter zero ou mais atributos. Para o *sujeito* e *objeto*, está-se interessado em extrair como atributos os seus *pré e pós-modificadores*, que podem ser *adjetivos*, *advérbios*, e *numerais* que os antecedem ou sucedem. Para o *verbo* está-se interessado em extrair os atributos indicados na Tabela 1.

Tabela 1. Possíveis atributos do verbo de uma relação (sujeito, verbo, objeto)

Atributo	**Descrição**
voz	Indica se o verbo está na voz *ativa* ou *passiva*.
negação	Indica se o verbo está na *afirmativa* ou *negativa*.
expressa	Indica se notícia tem uma pista explícita que indique que a ação expressa pelo verbo não aconteceu. Admitindo os valores: *"desejo de"*, *"possibilidade de"*. Por exemplo, do título *"Benfica pretende vencer a Liga"* é extraída a relação *(Benfica, vencer [expressa=desejo de], Liga)*

Para frases com mais de uma relação, o objetivo é extrair também a *inter-relação* entre cada duas relações adjacentes. Esta *inter-relação*, a existir, é obtida de conetores explícitos (ex. *mas, e, porque, porém*, etc.). O tipo de conetores (ex. *conetores de oposição*, *adição*, *confirmação*, etc.), indicam o tipo de interdependência entre as relações. Por exemplo, para a frase *"Real Madrid perde mas contínua na liderança e a praticar futebol muito forte",* o objetivo é extrair as seguintes *relações* R1:*(Real Madrid perde, null)*, R2:*(Real Madrid, continua, liderança)*, R3:*(Real Madrid, praticar, futebol [pos_mod1=muito, pos_mod2=forte])* e as *inter-relações* R4:*(R1, mas, R2)*, R5:*(R2, e, R3)*.

3 Trabalho Relacionado

A extração de relações (ER) tem sido estudada, por exemplo, em [4,10] para diversas finalidades. Por exemplo, para tarefas como resposta automática a perguntas [14], sumarização [15], alinhamento de ontologias [2], aprendizagem de ontologias [16], até tarefas mais amplas, tais como extração aberta de informação

(*Open Information Extraction*) [10,19]. Normalmente, estas tarefas não exigem ou parecem ser pouco afetadas pelos modificadores, o que pode ser a razão para não serem extraídos ou não lhes ser dada especial importância mesmo quando extraídos. O mesmo acontece com as *inter-relações* entre as relações extraídas.

Quanto à composição das relações extraídas, muitos dos estudos focam-se na extração de relações do tipo: *(arg1, relação, arg2)*. Por exemplo, existem trabalhos focados em relações que denotam aquisições de empresas, acontecimentos terroristas, trabalhos focados num conjunto pré-definido de relações, relações semânticas entre palavras ou conceitos [12], relações entre entidades mencionadas (ex.: pessoas, organizações, localizações, etc.) [5,11,21], ou em relações mais abrangentes como em [9,10,18]. Neste trabalho, tal como em [9,10] as relações são abrangentes e não recebem qualquer classificação semântica. O sistema ReVerb [9] reflete o atual estado da arte em termos de extração aberta de informação. Isto é conseguido usando e identificando os verbos através de um etiquetador gramatical (*part-of-speech tagger*). Por sua vez os argumentos da relação *(arg1 e arg2)* não estão restritos a entidades mencionadas. Portanto, a partir do título *"Real Madrid assume o primeiro lugar"*, é extraída a relação *(Real Madrid, assume, lugar [pre_mod=primeiro]),* que representa uma relação entre a entidade *"Real Madrid"* e o conceito *"primeiro lugar"*.

Várias abordagens têm sido aplicadas à ER. Desde abordagens que interpretam o problema de ER como um problema de classificação binária, aplicando em seguida um classificador baseado, por exemplo, em Máquinas de Suporte Vetorial [7,20], abordagens baseadas em modelos de máxima entropia [13], abordagens baseadas no caminho mais curto entre duas entidades num grafo de dependência [6], baseadas em regras [3,8,12]. Neste trabalho a ER baseia-se em regras. A falta de um corpus anotado e porque a criação de um seria uma tarefa demorada e dispendiosa, para além de apenas ser necessário um conjunto reduzido de regras, são razões que motivaram a abordagem aplicada.

Em conclusão, a extração de relações levadas a cabo em vários trabalhos, podem de alguma forma ser estendidas. Por exemplo, não ignorando ou dando especial atenção aos modificadores, não restringindo a extração a um conjunto pré-definido de relações, ou considerando as *inter-relações* entre relações extraídas. Este trabalho apresenta um estudo empírico preliminar, em que todas estas questões são levadas em conta. Embora algumas destas questões tenham sido abordadas separadamente, que os autores tenham conhecimento, este é o primeiro trabalho a combiná-las para a extração de relações.

4 News2Relations – Extração de Relações

Nesta secção é apresentado o sistema News2Relations 1.0, que está a ser desenvolvido para a extração de relações do tipo *(sujeito, verbo, objeto)*, respetivos *atributos* e *inter-relações*. O sistema é composto por vários módulos (Figura 1).

O sistema aceita como entrada títulos de notícias de desporto em Português, que passam por um conjunto de módulos, até se obter por fim um conjunto de relações. O funcionamento destes módulos é o seguinte:

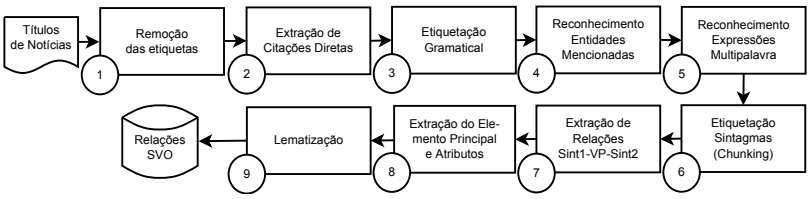

Figura 1. Módulos do News2Relations 1.0 para extração de relações.

1. **Pré-processamento: remoção das etiquetas.** Remove todas as etiquetas dos títulos que as contenham (Tabela 2). As etiquetas são todas as palavras que antecedem a primeira ocorrência de ": " ou "– ", desde que o título não seja uma citação direta. Neste último caso, as palavras só são removidas, se forem encontrados dois dos referidos símbolos. Estas etiquetas fornecem um contexto útil para os leitores humanos, mas podem afetar negativamente os passos seguintes. No conjunto de treino, 25% dos títulos contêm etiquetas.

Tabela 2. Exemplos de títulos de notícias e etiquetas a serem removidas

Título de notícia	**Etiqueta a remover**
Futsal Cup: Benfica vence Luparense (8-4) e está na final	"Futsal Cup: "
Diego: "Difícil perder desta maneira"	citação direta
"Real de Madrid tem de provar o que vale em campo" — Figo	citação direta
L. Europa – Benayoun: "Estou certo que podemos vencer..."	"L. Europa – "
Inglês disputado como comentador	sem etiqueta

2. **Pré-processamento: extração de citações.** Para os títulos que são citações diretas, é extraída a citação e o seu autor. Nos passos seguintes, é usada a citação extraída em vez de todo o título de notícia. No 2.º, 3.º e 4.º título da Tabela 2 é usada a frase entre aspas que correspondem às palavras ditas respetivamente por *Diego*, *Figo* e *Benayoun*.
3. **Etiquetação gramatical (part-of-speech tagging) e filtragem das frases.** Primeiro, cada palavra é anotada com a sua classe gramatical, tais como *substantivo*, *verbo*, *adjetivo*, *preposição*, etc. Depois, são ignoradas todas as frases sem qualquer verbo porque o objetivo é extrair relações *(sujeito, verbo, objeto)*, em que o verbo é obrigatório. No conjunto de dados analisado, 18% dos títulos não possuem qualquer verbo. Exemplos: "*Dakar: Hélder Rodrigues em terceiro*", "*Vitória tangencial do Benfica*".
   Esta tarefa é realizada com o etiquetador gramatical (*Part-of-Speech Tagger*) do OpenNLP[3] 1.5.2, usando um modelo de máxima entropia, treinado com o Bosque[4] 8.0, uma parte do corpus Floresta Sintá(c)tica [1] totalmente revista por linguistas.

[3] http://opennlp.sourceforge.net/
[4] http://www.linguateca.pt/floresta/corpus.html

4. **Reconhecimento e classificação de entidades mencionadas.** *Pessoas*, *organizações*, *locais*, *eventos*, etc., são reconhecidas e classificadas. Porém, o reconhecimento é suficiente para garantir que entidades compostas por mais de uma palavra sejam vistas como uma única unidade. O reconhecimento é feito por um sistema baseado em regras, previa e especialmente desenvolvido para o reconhecimento de entidades em notícias desportivas, pois contém uma série de pistas e listas (*gazetters*) próprias para este domínio.
5. **Reconhecimento de expressões multipalavra.** Neste módulo, são anotadas todas as expressões multipalavra contidas num dicionário existente para o domínio desporto. Estas são expressões, cujo seu significado é diferente do obtido pela análise individual das palavras que a compõem. Por exemplo, *"Pontapé de bicicleta"* é um tipo de remate em futebol, enquanto *"Pontapé"* significa chuto, e *"bicicleta"* é um veículo de duas rodas.
6. **Etiquetação dos sintagmas (Chunking)**. As sequências de palavras são anotadas em sintagmas não sobrepostos usando o *chunker* do OpenNLP 1.5.2, com um modelo de máxima entropia treinado no conjunto de treino referido no passo 3. Este produz as seguintes anotações: NP (sintagma nominal), VP (sintagma Verbal), PP (sintagma proposicional), ADJP (sintagma adjetival) e ADVP (sintagma adverbial). Por exemplo: "[NP *Inter*_n] [PP *de*_prp] [NP *Milão*_prop] [VP *vence*_v-fin] [NP *Livorno*_n] *e*_conj-c [VP *reforça*_v-fin] [NP *liderança*_n]".
   Se no passo 4 uma entidade mencionada é reconhecida ou se no passo 5 é reconhecida uma expressão multipalavra, que se sobrepõe a mais de um sintagma, estes são fundidos num único. Por exemplo, se no passo 4 for reconhecida a entidade "*Inter de Milão*", os 3 sintagmas (NP, PP, NP) do exemplo anterior são fundidos em "[NP *Inter de Milão*]...".
7. **Extração de relações (SINTAGMA1, VP, SINTAGMA2) e interrelações**. Para cada VP, extrai um sintagma como sujeito e outro como objeto. Para frases com dois ou mais VPs: I) extrai a interdependência entre cada duas relações adjacentes; II) infere o sujeito para a segunda relação, com base na relação precedente, se necessário e possível. Este módulo é explicado em maior detalhe na secção 5.2.
8. **Extração do elemento principal e atributos**. A partir de cada sintagma das relações *(SINTAGMA1, VP, SINTAGMA2)* é extraído o elemento principal e os seus atributos para obtenção das relações *(sujeito [atributos], verbo [atributos], objeto [atributos])*. Por exemplo, dado que um NP tem como palavra central o nome, este é extraído como o elemento principal da relação e os seus prés e pós-modificadores (ex. adjetivos) como atributos. Dado que um VP tem como palavra central o verbo, este é extraído como elemento principal da relação. No caso de um VP conter mais do que um verbo, assume-se que o verbo principal é sempre o último e por isso é o extraído (exemplo 1, na secção 1). Para extrair os atributos do verbo (*voz*, *negação*, e *expressa*), são usadas atualmente as seguintes heurísticas:
   - Está-se perante a voz passiva, se o verbo principal tiver sido anotado no passo 3 como *v-pcp* (verbo particípio). Exemplo: *"NKake <u>cedido</u> ao Sp. Covilhã"*. Esta heurística é imperfeita, pois o particípio também pode ser em-

pregue na voz ativa. Além do mais, a voz passiva, pode expressar-se de outras formas.

- Está-se perante uma negação, se existir uma expressão de negação (ex. *já mais*, *não*, *nunca*, *sem*, etc.) que anteceda imediatamente o sintagma verbal. No conjunto de treino, a negação ocorre em 5% dos títulos, sendo que esta heurística cobre 4% dos títulos.

- O verbo expressa uma ação que não aconteceu, quando existem explicitamente determinadas palavras (ex. *ambicionar*, *dever*, *sonhar*, etc.) que antecedem o verbo principal ou o sintagma verbal. Como por exemplo, no título *"Rússia quer conquistar 25 medalhas de ouro em Londres"*.

9. **Identificação da raiz das palavras flexionadas.** Neste passo, as palavras são reduzidos ao seu lema. Este procedimento tem como objetivo facilitar a comparação das palavras normalmente existentes em dicionários.

   Esta tarefa é realizada com o Tree-tagger [17], usando um modelo[5] para o Português, já treinado por Pablo Gamallo.

Este processo de extração de relação enfrenta uma série de desafios. **No passo 1** é necessário remover as etiquetas, evitando desta forma, por exemplo, que estas sejam erradamente identificadas como sujeito da relação. **No passo 2** é necessário extrair a citação e usá-la nos passos seguintes. Se este passo não fôr realizado, pode acontecer, por exemplo, o autor da citação ser incorretamente identificado como o sujeito ou objeto de uma relação. **No passo 3**, se o etiquetador atribuir incorretamente uma etiqueta gramatical, várias consequências podem ocorrer. O erro mais crítico ocorre quando uma palavra é detetada incorretamente como sendo verbo, pois conduz à extração de uma relação totalmente errada. Outro erro importante ocorre quando frases contendo verbos são ignoradas porque nenhum foi detetado. **No passo 4**, as abreviaturas que podem surgir (talvez para poupar espaço) podem conduzir a erros no reconhecimento de entidades mencionadas (por exemplo, *Man. United* em vez de *Manchester United*). **No passo 5**, o reconhecimento das expressões multipalavra depende da qualidade do dicionário. **No passo 6**, o etiquetador de sintagmas pode atribuir uma etiqueta errada. Semelhante ao passo 3, o erro mais crítico ocorre quando um VP não é detetado ou é detetado incorretamente. **No passo 7**, a maior dificuldade advém de títulos que são citações diretas, pois podem conter padrões muito diversificados. Estes padrões obrigam à definição de mais regras, tornando o sistema mais complexo. **No passo 8**, a extração da deteção da voz passiva é por si só complexa. **No passo 9**, a deteção dos lemas das formas verbais representa a maior dificuldade. Por exemplo, a forma verbal "*foi*" pode ter como lema "*ser*" ou "*ir*", dependendo do contexto.

Finalmente, porque há um erro associado com cada passo, a probabilidade da relação extraída conter erros, aumenta a cada passo.

[5] http://gramatica.usc.es/~gamallo/tagger.htm

5 Experiências

5.1 Conjunto de Dados de Treino

Primeiro, recolheram-se feeds RSS de notícias desportivas em Português com especial interesse no futebol. Em seguida, foi retirado aleatoriamente uma amostra de 200 títulos e realizada uma análise manual. Destes, 89% eram sobre ou relacionados ao futebol e os restantes 11% sobre outros desportos. Dos 200 títulos, 18% não continham qualquer verbo, 64% continham um, 16% continham dois, e 2% continham três verbos. Em média cada título continha 7 palavras, 43 caracteres.

Tabela 3. Frequência dos padrões do sujeito (SINTAGMA1) e objeto (SINTAGMA2), das relações (SINTAGMA1, VP, SINTAGMA2), encontrados no conjunto de treino. O texto sublinhado indica o sintagma a extrair. O texto a *itálico* o VP. Nas colunas da frequência como sujeito e como objeto, o texto que condiz com o padrão está a negrito.

id	Padrão	Frequência como Sujeito	Frequência como Objeto
1	**sem verbos**	18% Sucesso da seleção Inglesa de futebol crucial para o treinador	
2	NP	61% **Brasil** *goleia* Chile (4-0)	22% Jorge Jesus: "Só *podia haver* **um vencedor**"
3	Null	12% *Formalizado* acordo com a FIFA	6% Manchester United: Beckford *pretende sair*
4	NP (PP NP)+	6% "**Jogo com o Benfica** *é* para *vencer*" – Daniel Cruz	22% AC Milan *vence* (3-0) **Juventus com golos de Ronaldinho**
5	NP (,NP)? ,? e|ou NP	2% **Ben Gordon e Rodney Stuckey** *falham* jogo contra Boston Celtics	0%
6	PP NP	0%	17% "Aqui *sinto-me* **em casa**" - Cristiano Ronaldo
7	PP NP (PP NP)+	0%	5% Man. United: Cristiano Ronaldo *premiado* **com Golo da Década**
8	**Outros Padrões**	1%	10%

5.2 Extração de Relações (SINTAGMA1, VP, SINTAGMA2)

Nesta secção é apresentada a forma como a extração das relações *(SINTAGMA1, VP, SINTAGMA2)* está implementada no módulo 7 do sistema apresentado na secção 4. A extração é feita em duas fases. A primeira aplicada a todos os títulos de notícias com pelo menos um VP. A segunda a títulos com mais que um VP.

Na primeira fase existem 4 regras (Tabela 4). O objetivo é extrair um sintagma como sujeito e um outro como objeto para cada VP contido na frase. Estas regras usam os VPs como âncoras e resultaram da análise dos padrões do conjunto de treino (Tabela 3, secção 5.1). Para a extração do sujeito é aplicada a regra 1 ou 2, sendo escolhida sempre a que cobre mais texto. Para extração do

objeto igualmente mas aplicando a regra 3 ou 4. Nesta fase sintagmas entre dois VPs, podem ser identificados simultaneamente como sujeito de um VP e objeto de outro. Como um sintagma, se não sempre, na maioria dos casos é sujeito ou objeto, mas não ambos, é preciso resolver esta ambiguidade na 2.ª fase.

Tabela 4. Regras para extração de um sintagma sujeito (1 e 2) e um objeto (3 e 4) de um VP. Os sintagmas sublinhados são os extraídos se a regra é aplicada.

1	NP1 (PP NPn)* {negação}? **VP**			3	{negação}? **VP** (PP	ADVP)? NP	
2	NP1 (, NPn)*,? e	ou NPn+1 {negação}? **VP**		4	{negação}? **VP** (ADJP	ADVP)	

Numa segunda fase, um segundo conjunto de regras é aplicado (Tabela 5), mas somente para frases com mais de um VP. Nessas regras, dois VPs adjacentes são usados como âncoras. Os objetivos são: 1) extrair a inter-relação entre as duas relações adjacentes; 2) inferir um sujeito para a segunda relação, se necessário e possível, com base na primeira relação; 3) resolver possíveis ambiguidades causadas pela aplicação de regras na primeira fase, que identifiquem um sintagma simultaneamente como sujeito de uma relação e objeto de outra (exemplo 5).

Tabela 5. Exemplo de 3 das 6 regras aplicadas na 2.ª fase.

Id	Regra		Inter-relação	Sujeito de R2 herdado do	Necessário resolver possível ambiguidade
1	**VP1** (PP	ADVP)? NP? Conj **VP2**	(R1, Conj, R2)	sujeito de R1	não
2	**VP1** Conj NP **VP2**		(R1, Conj, R2)	não herda	não
3	**VP1** PP? NP1 Conj NP2 PP? **VP2**		(R1, Conj, R2)	não herda	sim**
(**Se esta regra é aplicada, é possível que a regra 2 e 3 da primeira fase também o tenham sido, criando uma ambiguidade. O título de exemplo 5, ilustra esta situação e respetiva solução.)					

Para ilustrar como as regras são aplicadas em ambas as fases, tome-se como exemplo os títulos de notícias seguintes. Os exemplos estão etiquetados com os sintagmas obtidos no passo 6 pelo *chunker* (secção 4) e as conjunções etiquetadas com o valor */Conj* no passo 3.

1. Euro 2012: [*NP1* Ucrânia], [*NP2* Suécia], [*NP3* França], e*/Conj* [*NP4* Inglaterra] [**VP** estão] [*PP* no] [*NP5* grupo D]
2. "[*ADVP* Não] [**VP** correu] [*ADVP* bem]" - Van der Gaag
3. [*NP1* Ajax] [**VP1** vence] e*/Conj* [**VP2** mantém] [*NP2* pressão] [*PP* sobre] [*NP3* o líder]
4. WRC: [*NP1* Loeb] [**VP1** vence] e*/Conj* [*NP2* Araújo] [**VP2** lidera]
5. Volta à Catalunha: [*NP1* Malacarne] [**VP1** vence] [*NP2* etapa] e*/Conj* [*NP3* Rodriguez] [**VP2** mantém] [*NP4* liderança]

Na primeira fase são aplicadas as regras da Tabela 4 para extração de um sujeito e um objeto para cada VP, obtendo-se as relações da Tabela 6.

Tabela 6. Aplicação das regras da 1.ª fase para extração de relações (SINTAGMA1, VP, SINTAGMA2)

Exemplo	Âncora	Regra Sujeito	Regra Objeto	Relações Extraídas
1	VP	2	3	(Ucrânia, estão, grupo D), ..., (Inglaterra, estão, grupo D)
2	VP	-	4	(null, correu [negação=sim], bem) (o atributo "negação", só é extraído no passo 8)
3	VP1	1	-	R1: (Ajax, vence, null)
	VP2	-	3	R2: (null, mantém, pressão)
4	VP1	1	-	R1: (Loeb, vence, null)
	VP2	1	-	R2: (Araújo, lidera, null)
5	VP1	1	3	R1: (Malacarne, vence, etapa)
	VP2	2	3	R2: (etapa e Rodriguez, mantém, liderança)

A Tabela 6, mostra, por exemplo, que para o título de exemplo 3, usando como âncora o VP1, é aplicada a regra 1 que extrai o NP1 (Ajax) como sujeito. Usando ainda o VP1 como âncora, não é aplicada nenhuma regra para extração do objeto, pois não existe objeto. Obtém-se então a relação R1:*(Ajax, vence, null)* para o VP1. Usando como âncora o VP2, nenhuma regra é aplicada para a extração do sujeito, mas é aplicada a regra 3 que extrai o NP2 (pressão) como objeto, obtendo-se a relação R2:*(null, mantém, pressão)*.

O título de exemplo 5, ilustra um caso de ambiguidade, porque *"etapa"* foi extraída como objeto de VP1 pela regra 3, mas também como parte do sujeito de VP2 pela regra 2. Como um sintagma é objeto de uma relação ou sujeito de outra, mas não ambos, isto precisa ser resolvido na segunda fase.

Na segunda fase são aplicadas as regras da Tabela 5, aos títulos de exemplo com mais de um VP, obtendo-se as relações *(SINTAGMA1, VP, SINTAGMA2)* finais e *inter-relações* indicadas na Tabela 7.

Tabela 7. Aplicação das regras da 2.ª fase

Ex.	Relações após a 1.ª fase	Regra 2.ª fase	Relações finais após a 2.ª fase	Inter-relação
3	R1:(Ajax, vence, null) R2:(null, mantém, pressão)	1	R1:(Ajax, vence, null) R2:(**Ajax**, mantém, pressão)	(R1, e, R2)
4	R1:(Loeb, vence, null) R2:(Araújo, lidera, null)	2	R1:(Loeb, vence, null) R2:(Araújo, lidera, null)	(R1, e, R2)
5	R1:(Malacarne, vence, etapa) R2:(etapa e Rodriguez, mantém, liderança)	3	R1:(Malacarne, vence, etapa) R2:(**Rodriguez**, mantém, liderança)	(R1, e, R2)

5.3 Extração de relações - Avaliação

Na avalidação foram usadas 100 frases nunca vistas anteriormente. O sistema apresentado na seção 4 foi aplicado e as extrações avaliadas manualmente. Os resultados são mostradas na (Tabela 8).

A Tabela 8 mostra, por exemplo, que das relações *(sujeito, verbo, objeto)* que eram possíveis extrair, 80% estavam totalmente corretas. As restantes 20% continham um ou mais elementos da relação, errados.

Tabela 8. Avaliação da extração das relações (sujeito, verbo, objeto)

	Sujeito	Verbo	Objeto
80%	certo	certo	certo
9%	errado	errado	errado
5%	certo	certo	errado
4%	errado	certo	certo
2%	errado	certo	errado

Motivo da Incorreta Extracção	
40%	Regra de extracção (Módulos 7 e 8)
28%	Entidades mencionada incorreta (Módulo 4)
28%	Categoria gramatical incorreta (Módulo 3)
4%	Sintagma incorreto (Módulo 6)
0%	Módulos 1 e 2
Os Módulos 5 e 9 não foram avaliados	

Da totalidade de erros ocorridos, 40% deles ocorreram devido à aplicação incorreta de uma regra. Isto é causado principalmente por títulos com padrões complexos e pouco frequente. O segundo erro mais frequente, representando 28% da totalidade dos erros, deveu-se à identificação errada de uma entidade mencionada. Por exemplo, a relação *(Tiago, emprestado, At)* extraída da frase *"Tiago emprestado ao At. Madrid"* contém um objeto errado (incompleto), causado pela abreviatura *"At."*. Isto acontece porque o módulo de reconhecimento de entidades aceita como entrada um texto (um conjunto de frases) e não uma única frase. Como consequência, perante um título de notícia contendo uma abreviatura, por vezes, segmenta-o erradamente, por julgar estar perante duas frases. Outro erro frequente, com 28% da totalidade dos erros, deu-se na etiquetação gramatical, causado principalmente por verbos não identificados, mas também por palavras erradamente identificado como verbo. Um exemplo do primeiro é a não identificação do verbo *"empata"* na frase *"Man Utd. empata em Birmingham (1-1)"*. O erro menos frequente, representando 4% da totalidade dos erros, aconteceu na fase de etiquetação dos sintagmas, de uma forma e com consequências semelhantes à fase de etiquetação gramatical.

No que respeita às inter-relações, no conjunto de treino eram 5 aquelas que poderiam ser extraídas. Destas foram extraídas com sucesso 3. Nas restantes 2, um dos erros aconteceu antes sequer que fosse possível aplicar uma regra para a extração da inter-relação (A Tabela 5 contém exemplos destas regras). O erro deveu-se ao não reconhecido de um dos dois VPs possíveis, que são fundamentais para a aplicação dessas regras. O outro erro ocorreu por não estar definida nenhuma regra que permitiria extrair a inter-relação.

Um aspeto que não foi quantificado, mas é possível tirar uma conclusão qualitativa é o fato das regras de extração aplicadas (seção 5.2), em frases curtas, cobrirem todas as palavras da frase. Isto significa que o sentido superficial de toda a frase é potencialmente extraído. Por exemplo, isto acontece ao extrair-se a relação *(Brasil, goleou, Chile)* de "*O Brasil goleou o Chile*". Em frases longas, existem palavras que não são cobertas pelas regras aplicadas. Isto significa que parte do sentido frase não é sequer extraído. No entanto, observou-se que mesmo nestas circunstâncias, a relação extraída tende a cobrir o conteúdo principal da frase. Por exemplo, da frase "*Ben Gordon e Rodney Stuckey falham jogo contra os Boston Celtics*" ao extrair-se a relação *(Ben Gordon e Rodney Stuckey, falham, jogo)* mas ignorada a porção "*contra o Boston Celtics*".

6 Conclusões e Trabalho Futuro

Este artigo descreve um sistema para extração de relação a partir de títulos de notícias desportivas escritas em Português. As relações extraídas representam conhecimento estruturado que pode ser de grande utilidade para várias tarefas para a língua Portuguesa. Como trabalho futuro pretende-se estudar se estas relações podem ser usadas com sucesso na classificação de títulos de notícias como: *positivos*, *negativos*, *ambos*, ou *neutros*.

Como os resultados e os desafios apontados na secção 5.3 demonstram, o sistema ainda pode ser melhorado significativamente. Em termos da extração dos elementos principais das relações *(sujeito, verbo, objeto)*, estas melhorias podem ser feitas ao nível das regras de extração das relações (secção 5.2), do reconhecimento de entidades mencionadas, da etiquetação gramatical (*part-of-speech tagging*), e ao nível da etiquetação dos sintagmas (*Chunking*). Em termos da extração dos atributos dos elementos principais, as melhorias podem ser feitas na deteção da negação mas sobretudo da deteção da voz ativa e passiva.

No futuro além de procurar melhorar o sistema, gostaríamos de definir uma métrica que permitisse ter uma perceção do conhecimento que não é extraído pelas relações. Seria também interessante adaptar e aplicar o sistema a títulos de notícia de outros domínios (exemplo: *negócios*, *politica*, *saúde*, *ciência*, etc.).

Agradecimentos

António Paulo Santos é financiado pela *Fundação para a Ciência e Tecnologia* (FCT), através da bolsa SFRH/BD/47551/2008.

Referências

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: um treebank para o português. In: Gonçalves, A., Correia, C.N. (eds.) Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001). pp. 533–545. Lisboa, Portugal (2001)
2. Beisswanger, E.: Exploiting Relation Extraction for Ontology Alignment. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) Learning. pp. 289–296. LNCS, Springer (2010)
3. Ben Abacha, A., Zweigenbaum, P.: Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of biomedical semantics 2(Suppl 5), S4 (2011)
4. Brin, S.: Extracting Patterns and Relations from the World Wide Web. In: WebDB 98 Selected papers from the International Workshop on The World Wide Web and Databases. vol. 1590, pp. 172—-183. Springer-Verlag (1999)
5. Brun, C., Hagège, C.: Semantically-Driven Extraction of Relations between Named Entities. Science pp. 35–46 (2009)
6. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT 05. pp. 724–731. No. October, Association for Computational Linguistics (2005)

7. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics ACL 04. vol. 4, pp. 423–es. Association for Computational Linguistics (2004)
8. Drury, B., Almeida, J.J.: Identification of fine grained feature based event and sentiment phrases from business news stories. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics - WIMS '11. p. 1. ACM Press, New York, New York, USA (2011)
9. Etzioni, O., Fader, A., Christensen, J.: Open Information Extraction: The Second Generation. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. pp. 3–10. AAAI (2011)
10. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. Network pp. 1535–1545 (2011)
11. Freitas, C., Santos, D., Mota, C., Oliveira, H.G., Carvalho, P.: Relation detection between named entities : report of a shared task. Computational Linguistics 28(June), 129–137 (2009)
12. Gonçalo Oliveira, H., Costa, H., Gomes, P.: Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In: Luís S Barbosa, M.P.C.E. (ed.) INFORUM 2010 Actas do II Simpósio de Informática. pp. 537–548. Universidade do Minho (2010)
13. Kambhatla, N.: Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 22. Association for Computational Linguistics (2004)
14. Lee, C., Hwang, Y.G., Jang, M.G.: Fine-grained named entity recognition and relation extraction for question answering. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07. p. 799. ACM Press, New York, New York, USA (2007)
15. Leskovec, J., Grobelnik, M., Milic-Frayling, N.: Learning sub-structures of document semantic graphs for document summarization. In: LinkKDD Workshop. pp. 133–138. Citeseer (2004)
16. Poesio, M., Barbu, E., Giuliano, C., Romano, L.: Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations. In: ECAI 2008 3rd Workshop on Ontology Learning and Population. pp. 1–5 (2008)
17. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. vol. 12 (1994)
18. Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S.: Unsupervised discovery of scenario-level patterns for Information Extraction. In: Proceedings of the sixth conference on Applied natural language processing -. pp. 282–289. Association for Computational Linguistics, Morristown, NJ, USA (2000)
19. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: TextRunner : Open Information Extraction on the Web. Computational Linguistics 42(April), 25–26 (2007)
20. Zhao, S., Grishman, R.: Extracting relations with integrated information using kernel methods. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics ACL 05. pp. 419–426. No. June, Association for Computational Linguistics (2005)
21. Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J.R.: StatSnowball: a statistical approach to extracting entity relationships. In: Proceedings of the 18th international conference on World wide web - WWW '09. WWW '09, vol. 56, pp. 101–110. ACM Press, New York, New York, USA (2009)